FROM LATENT LOOPS TO BETTER NAVIGATION: UNDERSTANDING DIFFUSION-BASED WORLD MODELS WITH COGNITIVE-MAP-INSPIRED DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based world models have recently demonstrated strong visual navigation performance without explicitly constructing maps, yet the latent mechanisms behind this capability remain unclear. We analyze diffusion-based navigation world models through the lens of cognitive map theory. Using persistent homology, we find that these models develop low-dimensional latent spatial structures during navigation generation, suggesting the emergence of cognitive-map-like representations. However, these structures are unstable across layers and long-horizon autoregressive rollouts, which may contribute to accumulated prediction drift. Motivated by this observation, we propose a brain-inspired framework that augments a frozen diffusion world model with a compositional cognition module. This module introduces structured, action-conditioned grid–place dynamics as an explicit spatial prior to stabilize latent spatial representations and improve long-horizon consistency. Experiments on real-world navigation datasets show improved prediction, navigation, and representation stability over strong diffusion-based baselines. Component ablations support the contribution of structured grid–place conditioning, while correlation analysis reveals an association between topological persistence and trajectory accuracy. The code is available at anonymous GitHub repository.

1 INTRODUCTION

Spatial navigation is a fundamental capability for embodied agents, requiring them to infer spatial structure, update their state from egocentric observations, and plan actions toward a goal. Existing visual navigation methods typically follow three paradigms: constructing explicit spatial or topological maps through learned representations Chaplot et al. (2020); Marza et al. (2023), directly generating action sequences with policies Shah et al. (2023); Sridhar et al. (2024), or forecasting future egocentric observations with world models. Recently, diffusion-based world models (DWMs) have emerged as a powerful paradigm for visual navigation, enabling look-ahead prediction and model-predictive planning from observation-action histories Bar et al. (2025). Despite their strong empirical performance, the latent mechanisms that support their long-horizon navigation ability remain insufficiently understood.

A natural hypothesis is that successful navigation world models may develop internal spatial representations analogous, at an abstract level, to cognitive maps in biological systems. In humans and other mammals, cognitive maps support localization, path integration, and goal-directed navigation by organizing spatial structure into internal neural representations. Neuroscience studies associate these functions with interactions between the medial entorhinal cortex (MEC) and hippocampus (HC), where grid cells provide periodic multi-scale spatial codes and place cells provide localized, context-dependent representations Jacobs et al. (2010); Barry & Bush (2012); Hafting et al. (2005); Stensola et al. (2012); Smith & Mizumori (2006); Moser et al. (2017). Motivated by this perspective, we ask whether diffusion-based navigation models exhibit navigation-relevant latent spatial organization, and whether explicitly stabilizing such organization can improve long-horizon prediction and planning.

We first analyze the internal representations of a diffusion-based navigation world model using persistent homology. Our analysis reveals that low-dimension latent spatial structures can emerge

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

during navigation generation, suggesting the presence of cognitive-map-like organization. However, these structures are not consistently maintained across layers or during autoregressive rollouts, indicating representational instability under long-horizon prediction. We further observe that weaker context–goal continuity is associated with larger prediction drift, suggesting that unstable latent spatial organization may contribute to long-horizon inconsistency.

Motivated by these observations, we propose a brain-inspired framework that augments a frozen diffusion-based world model with a compositional cognition module. The module introduces structured grid–place dynamics as an explicit action-conditioned spatial prior. The resulting model preserves the generative capability of the diffusion backbone while providing deterministic spatial transitions for long-horizon navigation.

Our main contributions are summarized as follows:

- Identify loop-like latent organization in diffusion-based navigation models and establish an empirical association between topological persistence and navigation accuracy.
- Introduce a unified, brain-inspired framework modeling navigation as **predictive imagination** and **compositional cognition**, utilizing action-conditioned grid–place dynamics for navigation in complex real-world environments.
- Enhance DWMs with grid–place dynamics, providing deterministic state transitions and episodic indexing, improves generation and navigation performance and representation stability over strong diffusion-based navigation baselines.

2 Related Work

2.1 Spatial navigation algorithm

Spatial navigation is a fundamental problem for embodied agents, requiring perception, state estimation, and planning Fu et al. (2022); Sridhar et al. (2024). A common setting is image-goal navigation, where the agent must reach a target location given only a goal image and egocentric observations Krantz et al. (2022). Pioneering frameworks, such as Neural SLAM methods Yu et al. (2019); Chaplot et al. (2020), combine learned perception with classical mapping but are constrained by explicit reconstruction and heuristic planning, limiting scalability and robustness in long-horizon tasks. In contrast, policy-based methods trained via behavior cloning directly generate action sequences and are highly reactive Shah et al. (2023); Sridhar et al. (2024), yet they lack explicit long-horizon reasoning and error correction.

Recent work has explored world-model-based navigation. Diffusion-based navigation world model Bar et al. (2025) generates future egocentric observations conditioned on past observations and actions, enabling look-ahead planning, but remain sensitive to long-horizon distributional shifts and error accumulation. Hierarchical memory-based methods mitigate long-term dependencies via autoregressive updates and explicit memory structures, at the cost of increased architectural and training complexity Dong et al. (2025).

2.2 Paradigms for Modeling HC–MEC Navigation Circuits

To address these challenges, neuroscience-inspired modeling offers a promising direction. Humans and mammals rely on grid cells and place cells for self-localization and navigation: grid cells encode spatial metrics, while place cells represent specific locations and context.

Motivated by hypotheses and neural evidence of continuous attractors in biological systems Khona & Fiete (2022); Gardner et al. (2022), geometric approaches model navigation via continuous attractors or low-dimensional manifolds. Continuous Attractor Networks provide interpretable and stable state updates Wu et al. (2016); Zou et al. (2024); Claudi et al. (2025), but often lack rich sensory grounding and episodic memory.

Path integration has also inspired sequence-based navigation models, which flexibly integrate perception Whittington et al. (2020); Wang et al. (2024); Rooke et al. (2024), but mainly emphasize latent state evolution rather than explicit structured generative dynamics. Lie-group formulations impose

principled geometric constraints, yet do not by themselves yield scalable end-to-end systems Gao et al. (2021); Xu et al. (2022; 2025).

Several recent methods further introduce brain-inspired circuits for memory and navigation, such as world models combining discrete latent states with CAN-like dynamics Peng et al. (2025), HC–MEC inspired recurrent loops for memory-guided planning Wang et al., or grid–hippocampal codes as memory scaffolds Chandra et al. (2025). However, these works primarily focus on structured state transitions or memory capacity, while leaving compositional cognitive states and long-horizon imagination less explored.

Overall, existing systems often trade off explicit geometric dynamics, sensory grounding, and scalable episodic memory. This motivates our integration of biologically inspired deterministic dynamics into diffusion-based navigation world models.

3 PRELIMINARIES

Task Definition. We consider diffusion-based visual navigation with a Conditional Diffusion Transformer (CDiT). At timestep t, an agent observes an egocentric RGB image $x_t \in \mathcal{X}$ and executes an action $a_t = (\mathbf{u}_t, \phi_t)$, where $\mathbf{u}_t \in \mathbb{R}^2$ denotes planar translation and ϕ_t denotes yaw rotation. A trajectory is

$$\mathcal{D}_{0:T} = \{x_0, (a_0, x_1), \ldots, (a_{T-1}, x_T)\}.$$

Given a context window $s_{t-\tau:t}$, an action sequence $a_{t\to t+k}$, and a prediction horizon k, the world model learns the k-step predictive distribution

$$p(s_{t+k} \mid s_{t-\tau:t},\, a_{t\to t+k},\, k)\,. \tag{1}$$

Distribution continuity. Diffusion-based navigation models predict future states by iterative denoising under contextual and action conditions. As the context–target discrepancy grows, the conditional future distribution becomes more multi-modal and uncertain, making long-horizon denoising more prone to drift and trajectory inconsistency. A detailed formulation is provided in Appendix B.

Cognitive-map prior. We use cognitive-map-like representations to refer to latent spatial structures that support localization, path integration, and goal-directed navigation Tolman (1948). Inspired by grid-cell and place-cell systems, where periodic multi-scale codes and localized contextual codes jointly support spatial reasoning Khona & Fiete (2022); Sorscher et al. (2019), we introduce structured grid-like dynamics as an explicit spatial prior. Additional biological background is provided in Appendix C.

4 INTERNAL REPRESENTATIONS IN DIFFUSION-BASED NAVIGATION MODELS

4.1 EMERGENCE OF LOOP-LIKE LATENT REPRESENTATIONS

To understand why diffusion-based navigation world models (NWM) perform well, we analyze a CDiT-XL model Bar et al. (2025) with 28 CDiT blocks and 1B parameters. Focusing on intermediate diffusion steps over a 16-second horizon (64 latent states) avoids incoherent early representations and late-stage collapse. We evaluate two settings: (i) generation, as action-conditioned direct generation, and (ii) navigation, as continuous autoregressive rollout.

Persistent homology reveals a prominent loop-like structure in the latent representations (Fig. 1, Appendix E.3). Its relationship to navigation is examined through context–target perturbation and the persistence–ATE analysis. This structure consistently emerges around block 10 in both settings.

Although classical grid and place cell theories predict a two-dimensional toroidal topology (T^2), The observed S^1 structure may reflect periodic variables coupled along the sampled navigation trajectories. This can be explained by the trajectory-centric training regime, the projection of latent dynamics into a reduced observation space, and the task's emphasis on goal-directed navigation along constrained paths rather than uniform 2D exploration. Together, these factors suggest that the observed loop-like structure reflects a task-adaptive abstraction of spatial structure rather than a deficiency in representational capacity. We also explored the geometry of the model's internal representations under different settings, with results provided in the Appendix E.6.

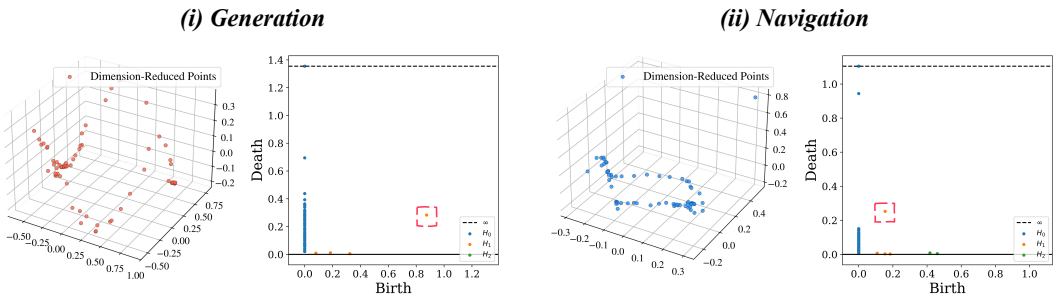

Figure 1: **Manifold and Topological Comparison.** Left: learned low-dimensional manifolds under (i) and (ii). Right: corresponding persistence diagrams highlighting the H1 circular feature.

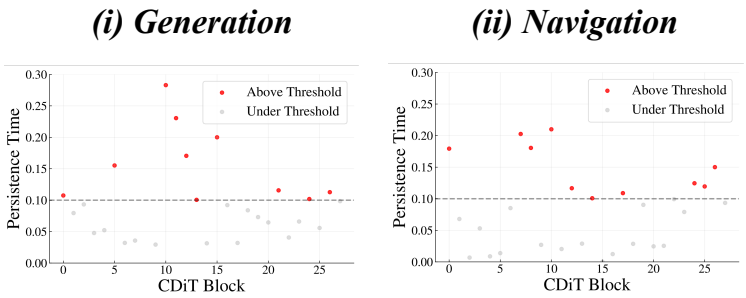

Figure 2: **Stability of Representations.** The lifetimes of prominent H1 features across layers are visualized, with significant features (> 0.1) highlighted.

4.2 Stability of Cognitive-Map–like Representations

To assess representation stability, we computed persistence diagrams across all layers and extracted the lifetimes of prominent H1 features (loops in the manifold). As shown in Fig. 2, significant cognitive-map-like representations appeared mainly in early-to-mid layers, present in roughly one-third of layers, indicating they are not consistently maintained. These features were more pronounced in the generation task than in navigation, likely due to error accumulation disrupting map-like structures. Statistical analyses over multiple samples are reported in Appendix E.7.

4.3 Cognitive-Map-Like Representations Reflect Spatial Structure

To further test whether the observed topology reflects navigation-relevant spatial structure, we perturb the spatial consistency of training data by injecting 25% mismatched context–target pairs and train the original NWM. For these negative samples, the true future target frame in the RECON dataset is replaced with an incorrect future frame while the context remains unchanged. This intervention tests whether the observed topology depends on consistent context–action–target relationships.

Empirically, introducing mismatched pairs leads to degraded performance even when the model is evaluated on the original correct test set. More importantly, the internal topological structure progressively weakens and eventually collapses in Fig. 3. These results suggest that the observed topology is tied to the spatial consistency required for navigation, supporting our interpretation that it reflects an organized cognitive-map-like structure rather than a byproduct of visual statistics.

5 Enhancing DWMs via Cognitive Maps

The analysis above links latent topological organization to navigation performance, motivating explicit action-conditioned grid–place dynamics as a spatial prior for diffusion world models. To improve

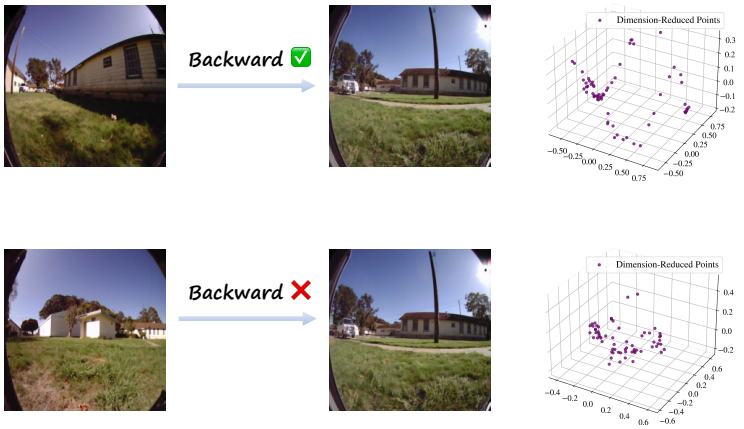

Figure 3: **Comparison of Positive and Negative Samples.** Context–target mismatch leads to the collapse of cognitive-map-like representations.

the temporal stability of spatial prediction, we introduce a cognitive-map-conditioned diffusion framework that augments the denoising process with explicit action-conditioned spatial dynamics.

Our model consists of two components. First, a frozen CDiT backbone serves as the **predictive imagination module**, which preserves the pretrained generative ability of the diffusion world model. Second, a trainable **compositional cognition module** produces a structured spatial condition through grid–place dynamics. The overall architecture is shown in Fig. 4(c).

Predictive Imagination Module. The predictive imagination module models future states using a frozen CDiT backbone. Given a context window $s_{t-\tau:t}$, a cumulative action $a_{t\to h}$, a relative prediction horizon k, and a cognitive state c_h, it predicts the future state distribution as

$$s_h \sim P_\theta(s_h \mid s_{t-\tau:t}, a_{t\to h}, k, c_h). \tag{2}$$

Here, c_h is produced by the compositional cognition module and serves as a structured spatial condition for denoising.

Following NWM, we represent actions as cumulative continuous displacements over the prediction window. The action from s_t to s_h is defined as

$$a_{t\to h} = (\mathbf{u}_{t\to h}, \phi_{t\to h}), \quad \mathbf{u}_{t\to h} = \sum_{i=t}^{h} \mathbf{u}_i, \quad \phi_{t\to h} = \sum_{i=t}^{h} \phi_i \ (\mathrm{mod}\ 2\pi). \tag{3}$$

This cumulative action representation provides a smoother conditioning signal for long-horizon prediction than using isolated step-wise actions.

Compositional Cognition Module. To reduce long-horizon drift, we introduce a compositional cognition module that provides deterministic, action-conditioned spatial dynamics. Inspired by multi-scale grid representations in biological navigation, the module represents spatial cognition using a compositional grid code

$$g = [g^{(1)}, g^{(2)}, \ldots, g^{(M)}], \tag{4}$$

where each grid module $g^{(i)}$ operates at a distinct spatial period λ_i. This multi-module structure enables spatial states to be encoded at multiple resolutions.

Given a sensory history $s_{t-\tau:t}$, an encoder first produces a place-like latent representation:

$$p_t = \mathrm{Encoder}(s_{t-\tau:t}). \tag{5}$$

The place representation is projected into grid space, evolved by deterministic path integration, and then projected back to place space:

$$g_t = \mathcal{P}(p_t), \qquad g_h = T(a_{t\to h})g_t, \qquad p_h = \mathcal{Q}(g_h). \tag{6}$$

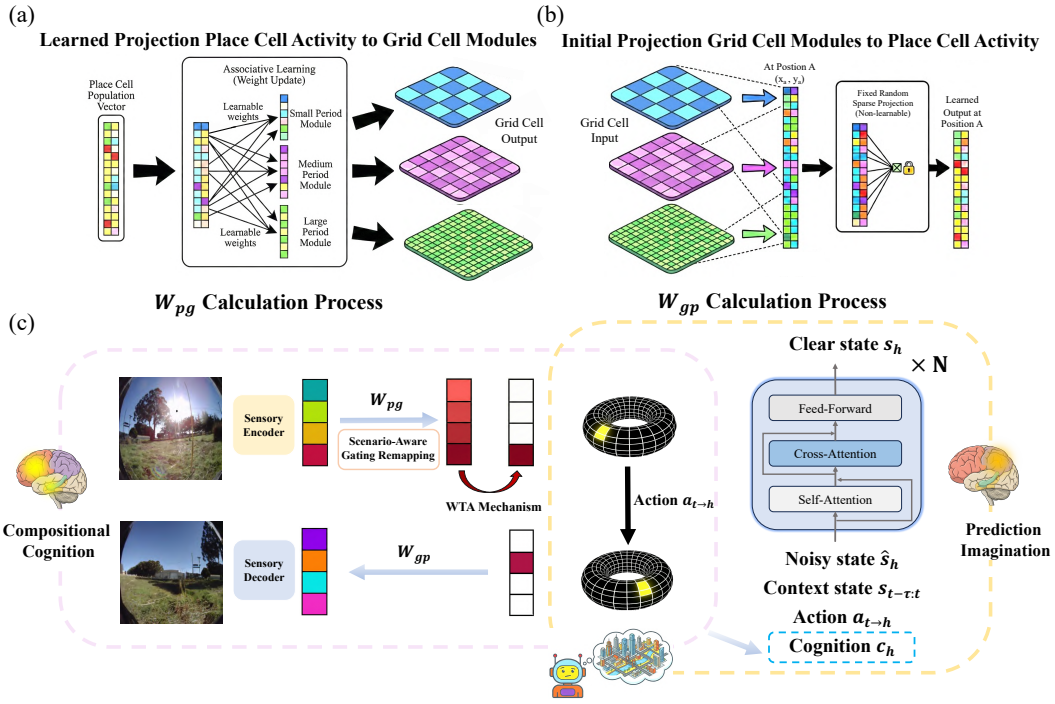

Figure 4: **Model Architecture.** Panels (a) and (b) show the grid-to-place and place-to-grid projection processes, respectively. Panel (c) illustrates the cooperation between the frozen **predictive imagination module** and the trainable **compositional cognition module**.

Here, $\mathcal{P}$ denotes the place-to-grid projection, $\mathcal{Q}$ denotes the grid-to-place projection, and $T(\cdot)$ is an action-conditioned transition operator applied to the grid modules.

The evolved place representation p_h is used to form the cognitive condition for the diffusion backbone:

$$c_h = k \cdot p_h. \tag{7}$$

This condition couples the relative prediction horizon with the post-action spatial state, providing the frozen CDiT backbone with an explicit spatial prior during denoising.

Importantly, this module is used as a biologically inspired inductive bias rather than as a direct simulation of the medial entorhinal cortex–hippocampus circuit. The deterministic grid transition encourages temporal coherence, while the diffusion backbone retains its ability to model uncertainty and generate future observations. Details of the neuroscientific motivation, method implementation details, and representational capacity analysis are provided in Appendix A.

Optimization. The CDiT backbone is frozen during training, and only the **compositional cognition module** and its conditioning interface are optimized. The main denoising loss supervises whether the cognitive state provides useful spatial conditioning for future-state prediction:

$$\mathcal{L}_{\text{imagination}} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(s_h \mid s_{t-\tau:t}, a_{t\to h}, k, c_h)\|_2^2\right]. \tag{8}$$

We further regularize the cognition module with a future-state reconstruction loss and a zero-action anchoring loss. The overall objective is

$$\mathcal{L} = a_1\mathcal{L}_{\text{imagination}} + a_2\mathcal{L}_{\text{reconstruct}} + a_3\mathcal{L}_{\text{anchor}}, \tag{9}$$

where a_1, a_2, and a_3 are weighting coefficients. Definitions and implementation details of the training objectives are provided in Appendix A.7.

Table 1: **Image/Video Generation Performance Comparison.** 16s prediction evaluation and video generation on the RECON dataset.

Method	DIAMOND	NWM	Ours
DreamSim ↓	0.255 ± 0.001	0.176 ± 0.001	**0.169** ± 0.001
LPIPS ↓	0.628 ± 0.004	0.431 ± 0.002	**0.401** ± 0.002
PSNR ↑	8.814 ± 0.015	12.997 ± 0.023	**13.301** ± 0.027
FID ↓	171.891 ± 4.216	31.226 ± 0.698	**28.979** ± 0.655
FVD ↓	725.549 ± 3.513	191.096 ± 5.237	**169.855** ± 5.011

Table 2: **Navigation Performance Comparison.** Navigation performance (ATE, RPE) evaluation on the RECON dataset.

Method	GNM	NoMad	NWM	Ours
ATE ↓	1.912 ± 0.015	1.875 ± 0.026	1.057 ± 0.017	**0.959** ± 0.014
RPE ↓	0.723 ± 0.009	0.509 ± 0.012	0.300 ± 0.010	**0.267** ± 0.011

6 EXPERIMENT

6.1 EXPERIMENTAL SETUP

Dataset. We train and evaluate on real-world robotics navigation datasets, including RECON Shah et al. (2021), SCAND Karnan et al. (2022), HuRoN Hirose et al. (2023), and TartanDrive Triest et al. (2022), and use Go Stanford Hirose et al. (2018) only for unseen-environment testing. Following prior work Bar et al. (2025); Dong et al. (2025), we compute relative actions from robot poses, rescale per-step translations by each dataset's average step length, and discard trajectories with backward motion or fewer than three steps.

Evaluation Metrics. We evaluate visual prediction with PSNR Hore & Ziou (2010), LPIPS Zhang et al. (2018), DreamSim Fu et al. (2023), FID Heusel et al. (2017), and FVD Unterthiner et al. (2019), and evaluate navigation with Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) Sturm et al. (2012). Additional metric details are provided in Appendix E.2. After training, we sample 5 times from each model to report mean and std results.

Implementation Details. We use a frozen NWM CDiT-XL as the predictive imagination module. The compositional cognition module uses grid periods $\lambda_i \in \{3, 4, 5, 7\}$ and $R = 18$ grid-to-place projections. We set $(a_1, a_2, a_3) = (4, 3, 0.5)$ and optimize trainable components with AdamW at learning rate 8×10^{-5}. Training uses 4 NVIDIA A100 GPUs with batch size 16. All models contain approximately 1B parameters.

6.2 OVERALL RESULTS

We conduct a systematic comparison between our method and competing models along two task dimensions: (i) the quality of action visual prediction (including image and video generation), and (ii) navigation performance.

The quality of action visual prediction. We compare our model with two DWMs, DIAMOND Alonso et al. (2024) and NWM. Within this class, our approach achieves state-of-the-art performance on image generation. Table 1 reports results on the RECON dataset at the 16s prediction horizon and video generation, showing that our advantage becomes more pronounced. Results on additional datasets are deferred to the Appendix E.9. Visualization results are shown in Fig. 5.

Navigation performance. For navigation, we adopt the MPC framework of NWM and compare against GNM Shah et al. (2022), NoMad Sridhar et al. (2024), and NWM. Table 2 reports results on the RECON dataset, where our method improves both navigation accuracy and pose consistency.

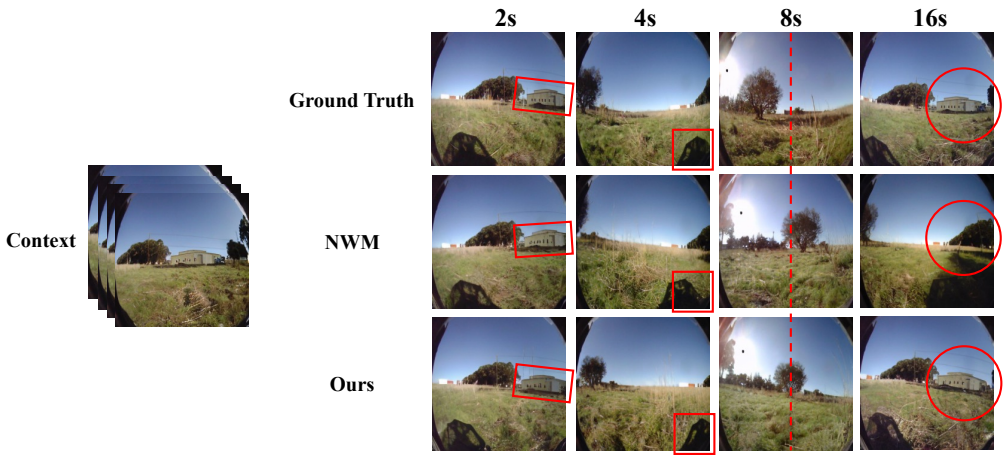

Figure 5: **Generation Performance.** The figure visualizes action generation results over 2 s, 4 s, 8 s and 16 s horizons on the RECON dataset. Red markers and dashed outlines indicate representative advantages of our approach and limitations of the baseline.

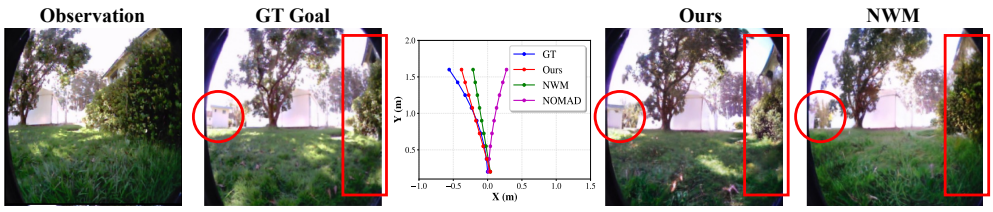

Figure 6: **Navigation Performance.** The figure reports navigation performance on the RECON dataset.

Table 3: **Generalization Performance Comparison.** 4s prediction evaluation on the Go Stanford dataset.

Method	NWM	Ours
DreamSim ↓	0.475 ± 0.001	**0.468** ± 0.001
LPIPS ↓	0.662 ± 0.002	**0.655** ± 0.002
PSNR ↑	11.029 ± 0.041	**11.497** ± 0.037

Results on additional datasets are deferred to the Appendix E.9. Visualization results are shown in Fig. 6.

Generalization in unseen environments. We evaluate 4-second generalization performance in the unseen Go Stanford environment (Table 3) and observe that our method improves the original model's generalization ability.

6.3 Ablation

Correlation Between Topological Persistence and Navigation Performance. To examine whether the detected topology is associated with downstream navigation behavior, we analyze 30 sampled trajectories from RECON (Fig. 7). We observe a negative correlation between topological persistence and ATE: trajectories with stronger persistent H_1 features tend to exhibit lower navigation error. This suggests that stable cognitive-map-like topology is associated with more coherent latent spatial organization and reduced long-horizon trajectory drift. Thus, persistent homology provides not only an interpretability tool, but also a quantitative lens for relating representation geometry to navigation reliability.

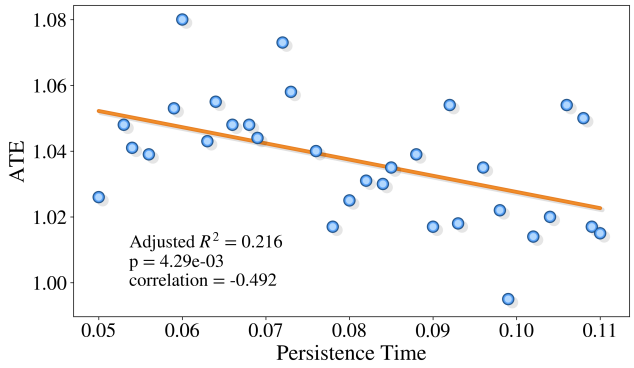

Figure 7: **Correlation Between Topological Persistence and Navigation Performance.** A stronger persistent H_1 feature is associated with lower ATE.

Table 4: **Component Ablation.** Component-wise ablation of the compositional cognition module on RECON.

Method	DreamSim ↓	LPIPS ↓	ATE ↓	RPE ↓
NWM backbone	0.176± 0.001	0.431± 0.002	1.057± 0.017	0.300± 0.010
MLP condition	0.175± 0.001	0.431± 0.002	1.059± 0.015	0.297 ± 0.009
w/ identity transition	0.175± 0.001	0.429± 0.002	1.023± 0.017	0.301 ± 0.010
w/o WTA	0.180± 0.001	0.442± 0.002	1.078 ± 0.021	0.312 ± 0.012
w/o remapping	0.173± 0.001	0.420± 0.002	1.016 ± 0.015	0.295 ± 0.010
Ours	**0.169**± 0.001	**0.401**± 0.002	**0.959**± 0.014	**0.267**± 0.011

Contribution of each cognitive-map component. To verify that the gains do not simply come from additional conditioning capacity, we compare our method with a matched MLP conditioning baseline and several component ablations. The MLP baseline uses the same frozen CDiT backbone and conditioning interface, but replaces the structured grid–place dynamics with a generic MLP. Other variants remove action-conditioned transition, WTA discretization, or scenario-aware remapping.

As shown in Table 4, the MLP condition performs almost on par with the NWM backbone, while our full model achieves clear improvements. Removing WTA degrades performance, and replacing path integration with identity transition or disabling remapping also weakens navigation accuracy. These results suggest that the gains mainly arise from structured grid–place dynamics rather than generic conditioning capacity.

Additional representation analyses. Appendix E.4 shows that (i) cognitive-map-like representations persist in unseen environments, and (ii) our method increases their persistence and stability in both generation and navigation settings. These results further support that the proposed cognitive-map prior stabilizes navigation-relevant latent organization beyond task-level metric improvements.

7 Discussion

Our results suggest that diffusion-based navigation world models develop navigation-relevant latent spatial structures, but these structures are unstable under long-horizon rollout. By introducing structured grid–place dynamics as an action-conditioned spatial prior, our method stabilizes these representations and improves both visual prediction and navigation accuracy. The current design uses an external cognitive-map module for clarity and modularity; future work could integrate compositional spatial structure directly into the diffusion backbone and scale the framework to larger multimodal world models.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Caswell Barry and Daniel Bush. From a to z: a potential role for grid cells in spatial navigation. *Neural systems & circuits*, 2(1):6, 2012.

Sarthak Chandra, Sugandha Sharma, Rishidev Chaudhuri, and Ila Fiete. Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, 638(8051):739–751, 2025.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.

Frédéric Chazal, Vin De Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

Federico Claudi, Sarthak Chandra, and Ila R Fiete. A theory and recipe to construct general and biologically plausible integrating continuous attractor neural networks. *bioRxiv*, pp. 2025–05, 2025.

Yifei Dong, Fengyi Wu, Guangyu Chen, Zhi-Qi Cheng, Qiyu Hu, Yuxuan Zhou, Jingdong Sun, Jun-Yan He, Qi Dai, and Alexander G Hauptmann. Unified world models: Memory-augmented planning and foresight for visual navigation. *arXiv preprint arXiv:2510.08713*, 2025.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Zipeng Fu, Ashish Kumar, Ananye Agarwal, Haozhi Qi, Jitendra Malik, and Deepak Pathak. 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.

Ruiqi Gao, Jianwen Xie, Xue-Xin Wei, Song-Chun Zhu, and Ying Nian Wu. On path integration of grid cells: group representation and isotropic scaling. *Advances in Neural Information Processing Systems*, 34:28623–28635, 2021.

Richard J Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A Baas, Benjamin A Dunn, May-Britt Moser, and Edvard I Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, 2022.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Noriaki Hirose, Amir Sadeghian, Marynel Vázquez, Patrick Goebel, and Silvio Savarese. 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.

Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Joshua Jacobs, Michael J Kahana, Arne D Ekstrom, Matthew V Mollison, and Itzhak Fried. A sense of direction in human entorhinal cortex. *Proceedings of the National Academy of Sciences*, 107 (14):6487–6492, 2010.

Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4): 11807–11814, 2022.

Mikail Khona and Ila R Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, 2022.

Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-specific image goal navigation: Training embodied agents to find object instances. *arXiv preprint arXiv:2211.15876*, 2022.

Pierre Marza, Laetitia Matignon, Olivier Simonin, and Christian Wolf. Multi-object navigation with dynamically learned neural implicit representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11004–11015, 2023.

Edvard I Moser, May-Britt Moser, and Bruce L McNaughton. Spatial representation in the hippocampal formation: a history. *Nature neuroscience*, 20(11):1448–1464, 2017.

Ji-Hoon Park, Yeong-Joon Ju, and Seong-Whan Lee. Explaining generative diffusion models via visual analysis for interpretable decision-making process. *Expert Systems with Applications*, 248: 123231, 2024.

Xiangyuan Peng, Xingsi Dong, and Si Wu. Vector quantization in the brain: Grid-like codes in world models. *arXiv preprint arXiv:2510.16039*, 2025.

Spencer Rooke, Zhaoze Wang, Ronald W Di Tullio, and Vijay Balasubramanian. Trading place for space: Increasing location resolution reduces contextual capacity in hippocampal codes. *Advances in Neural Information Processing Systems*, 37:96892–96923, 2024.

Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021.

Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. *arXiv preprint arXiv:2210.03370*, 2022.

Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023.

David M Smith and Sheri JY Mizumori. Hippocampal place cells, context, and episodic memory. *Hippocampus*, 16(9):716–729, 2006.

Ben Sorscher, Gabriel Mel, Surya Ganguli, and Samuel Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in neural information processing systems*, 32, 2019.

Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 63–70. IEEE, 2024.

Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, 2012.

Jürgen Sturm, Wolfram Burgard, and Daniel Cremers. 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/RJS International Conference on Intelligent Robot Systems (IROS)*, volume 13, pp. 6, 2012.

Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

Samuel Triest, Matthew Sivaprakasam, Sean J Wang, Wenshan Wang, Aaron M Johnson, and Sebastian Scherer. 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.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Zhaoze Wang, Genela Morris, Dori Derdikman, Pratik Chaudhari, and Vijay Balasubramanian. Remi: Reconstructing episodic memory during internally driven path planning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhaoze Wang, Ronald W Di Tullio, Spencer Rooke, and Vijay Balasubramanian. Time makes space: Emergence of place fields in networks encoding temporally continuous sensory experiences. *Advances in Neural Information Processing Systems*, 37:37836–37864, 2024.

James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

Si Wu, KY Michael Wong, CC Alan Fung, Yuanyuan Mi, and Wenhao Zhang. Continuous attractor neural networks: candidate of a canonical model for neural information representation. *F1000Research*, 5:F1000–Faculty, 2016.

Dehong Xu, Ruiqi Gao, Wen-Hao Zhang, Xue-Xin Wei, and Ying Nian Wu. Conformal isometry of lie group representation in recurrent network of grid cells. *arXiv preprint arXiv:2210.02684*, 2022.

Maoshen Xu, Fei Song, Yuxiu Shao, Bailu Si, and Shanshan Qin. The principle of isomorphism: A theory of population activity in grid cells and beyond. *arXiv preprint arXiv:2510.02853*, 2025.

Fangwen Yu, Jianga Shang, Youjian Hu, and Michael Milford. Neuroslam: A brain-inspired slam system for 3d environments. *Biological cybernetics*, 113(5):515–545, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

Xiaolong Zou, Xingxing Cao, Xiaojiao Yang, and Bo Hong. Leveraging attractor dynamics in spatial navigation for better language parsing. In *Forty-first International Conference on Machine Learning*, 2024.

A Compositional Cognition: Implementation and Modeling Details

A.1 Neuroscientific Motivation

Our compositional cognition module is inspired by the interaction between grid cells in the medial entorhinal cortex and place cells in the hippocampus. Grid cells exhibit periodic firing patterns at multiple spatial scales and are often modeled as providing a structured code for displacement and path integration. Place cells, in contrast, fire in localized regions of an environment and provide context-dependent spatial anchoring.

We do not aim to directly reproduce the biological MEC–HC circuit. Instead, we use this principle as an architectural inductive bias: the model maintains a structured latent spatial state and evolves it through action-conditioned dynamics.

A.2 Discrete Multi-Module Grid Representation

Each grid module $g^{(i)}$ is modeled as a discrete two-dimensional toroidal lattice with spatial period λ_i along both axes. The module is represented as a binary vector of length λ_i^2, and the full grid representation is obtained by concatenating all modules:

$$g = [g^{(1)}, g^{(2)}, \ldots, g^{(M)}], \quad g \in \{0,1\}^{N_G}, \tag{10}$$

where the total number of grid cells is

$$N_G = \sum_{i=1}^{M} \lambda_i^2. \tag{11}$$

This discrete multi-module design explicitly encodes periodic structure and supports compositional spatial representations across multiple scales.

A.3 Projection calculation in HC-MEC loop and scenario-aware gating mechanism

We model the interaction between grid cells and place cells using a bidirectional projection. The grid-to-place projection W_{gp} is randomly initialized and kept fixed throughout training. To enforce sparsity, a random mask with strength z is applied to W_{gp}. In contrast, the place-to-grid projection W_{pg} is initialized via associative learning and optimized during training. This asymmetric design induces a recurrent HC–MEC loop that serves as a high-capacity memory scaffold when W_{pg} is fixed Chandra et al. (2025). While W_{pg} is made learnable and updated during training, the network progressively acquires the ability to encode state transitions and store distinct memories.

A.4 Place-to-Grid Projections and Action-Driven Path Integration

Grid–to-place and place–to-grid projections are constructed to form a recurrent HC–MEC loop. Given a place cell representation p_t, the grid representation is obtained as:

$$g_t = WTA(W_{pg} \cdot p_t), \tag{12}$$

where $WTA(\cdot)$ denotes a winner-take-all operation applied independently within each grid module to preserve attractor-like dynamics.

State transitions in grid space are driven explicitly by physical actions through path integration. We construct an action-conditioned block-diagonal transition operator:

$$g_h = T(a_{t\to h}) \cdot g_t, \tag{13}$$

where each block corresponds to a grid module and implements translation on the associated toroidal lattice according to the action $a_{t\to h}$.

A.5 Grid-to-Place Decoding and Perceptual Interface

The predicted grid state is mapped back to the hippocampal space as:

$$p_h = ReLU(\tilde{W}_{gp} \cdot g_h), \tag{14}$$

where the ReLU nonlinearity enforces non-negativity of neural activations.

To connect compositional cognition with perception, we employ an encoder–decoder architecture:

$$p_t = Encoder(s_{t-\tau:t}), \quad \hat{s}_h = Decoder(p_h), \tag{15}$$

where the encoder maps a short temporal window of sensory observations to hippocampal representations, and the decoder reconstructs future observations from predicted states.

A.6 Scenario-Aware Gating Remapping Mechanism

To support multiple environments or contexts, we do not assume a single fixed grid-to-place projection (W_{gp}). Instead, we maintain a set of grid-to-place matrices $\{W_{gp}^{(k)}\}_{k=1}^{R}$, each randomly initialized, where R represents the number of candidate remappings, and use a learned gating network to produce context-dependent combination weights.

$$\tilde{W}_{gp} = \sum_{r=1}^{R} \alpha_r(s_t) W_{gp}^{(r)}, \quad \sum_r \alpha_r = 1. \tag{16}$$

This mechanism allows the model to flexibly reinterpret grid representations under changing environmental conditions, providing an abstract and functional account of the grid–hippocampal remapping phenomenon observed in biological systems.

A.7 Training Objectives

The cognition module is regularized by a future-state reconstruction loss:

$$\mathcal{L}_{\text{reconstruct}} = \frac{1}{D} \left\| \hat{s}_h - s_h \right\|_2^2, \tag{17}$$

where D denotes the sensory feature dimension.

When the action input is zero, the grid–place loop should preserve the current place representation. We therefore align the reconstructed place representation

$$\hat{p}_t = \text{ReLU}(\tilde{W}_{gp} g_t) \tag{18}$$

with the encoded place representation p_t using cosine similarity:

$$\mathcal{L}_{\text{anchor}} = 1 - \frac{\hat{p}_t^\top p_t}{\|\hat{p}_t\|_2 \|p_t\|_2}. \tag{19}$$

This loss stabilizes the place–grid–place loop and prevents unnecessary drift when no movement is applied.

A.8 Representational Capacity

The compositional grid code provides a compact but high-capacity spatial representation. With M modules of period λ, it can represent approximately λ^{2M} spatial states using only $M\lambda^2$ grid units. This factorized structure allows the model to represent large spatial domains while keeping the number of latent units manageable.

B Distribution Continuity in Diffusion Navigation

Diffusion-based navigation models generate future states by iteratively denoising under conditional inputs. When the target state is close to the observed context, the conditional future distribution is

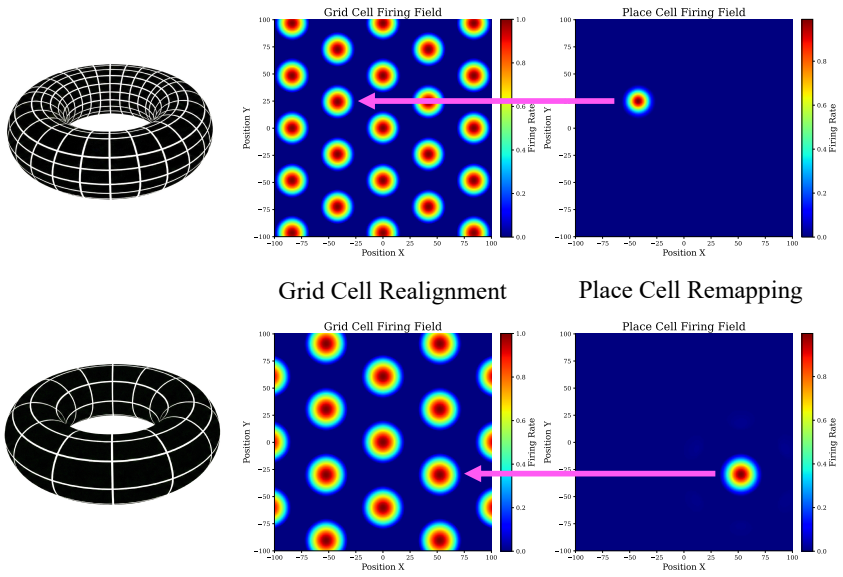

Figure 8: **Grid and Place Cell Representations.** Grid cell periodic firing induces a toroidal manifold structure, while transitions across environments illustrate grid phase realignment and corresponding place cell remapping.

relatively concentrated and the reverse diffusion process can follow a coherent trajectory. As the discrepancy between the context and the future target increases, however, the number of plausible future trajectories grows, leading to a more heterogeneous and higher-variance conditional distribution.

Let $d(s_{t-\tau:t}, s_{t+k})$ denote a discrepancy measure between the context and the target state. As this discrepancy increases, the conditional distribution may become multi-modal. We approximate it as

$$p(s_{t+k} \mid o) \approx \sum_{i=1}^{\mathcal{K}} w_i \mathcal{N}\big(s_{t+k}; \mu_i(o), \sigma_i^2(o)\big)\,, \tag{20}$$

where o denotes the context and conditioning variables, and $\mathcal{K}$ is the number of plausible future modes. A larger context–target discrepancy generally increases the variance of this conditional distribution:

$$\mathrm{Var}(s_{t+k} \mid o) \text{ tends to increase with } d(s_{t-\tau:t}, s_{t+k}).$$

This higher variance can induce rapidly changing or high-curvature conditional score fields, making long-horizon denoising more prone to distributional drift and trajectory inconsistency.

C Additional Details on Cognitive Map Representations and Path Integration

Fig. 8 illustrates the multi-scale spatial representations associated with grid cells and place cells. We provide a more detailed explanation below.

- **Grid cells:** Grid cells in the medial entorhinal cortex (MEC) fire at multiple locations that form a regular, hexagonal-like pattern across an environment. Cells within the same grid module share a similar spatial period and orientation but differ in phase, whereas different modules operate at different spatial scales. At the population level, these periodic firing patterns are often modeled as low-dimensional continuous attractors, such as toroidal manifolds, which provide a compact and structured code for spatial displacement.
- **Place cells:** Place cells in the hippocampus (HC) typically fire in one or a few localized regions of an environment. Compared with the periodic code of grid cells, place-cell activity provides a more localized and context-dependent representation of position. Interactions between MEC grid cells and hippocampal place cells are thought to support stable cognitive

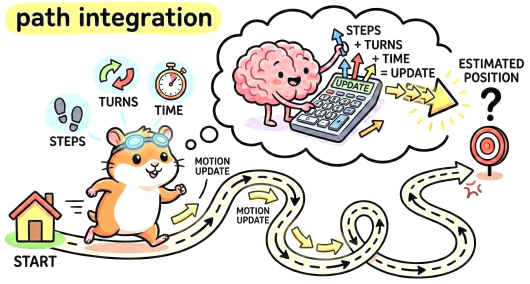

Figure 9: **Path Integration Schematic.** Position estimates are updated by accumulating self-motion cues, such as velocity and rotation.

maps by combining metric information about displacement with environment-specific location coding.

- **Phase realignment and remapping:** When an animal enters a new environment or when sensory context changes, grid-cell phases can shift and place-cell firing fields can remap. These changes do not imply a complete loss of spatial structure; rather, they allow the spatial code to be re-anchored to the current context. In this way, the grid–place cell system can support both stable navigation within an environment and flexible adaptation across different environments.

Path integration is the process by which an animal estimates its current position by accumulating self-motion cues over time. These cues may include linear velocity, angular velocity, displacement, or head-direction signals. This mechanism allows an animal to keep track of its location even when external landmarks are temporarily unavailable. In the mammalian brain, path integration is closely associated with the medial entorhinal cortex (MEC) and hippocampus (HC). Grid cells in MEC provide periodic spatial codes that are well suited for representing movement through space, while hippocampal place cells provide localized, context-dependent representations that can help anchor spatial estimates to a particular environment.

Formally, let $\mathbf{x}_t$ denote the agent's internal estimate of position at time t, and let $\mathbf{u}_t$ denote the self-motion signal. A simple discrete-time path integration update can be written as

$$\mathbf{x}_{t+1} = \mathbf{x}_t + F(\mathbf{u}_t) + \boldsymbol{\epsilon}_t,$$

where $F(\cdot)$ maps self-motion input to a displacement in the internal spatial representation, and $\boldsymbol{\epsilon}_t$ denotes accumulated integration noise or model error. Intuitively, this equation says that the agent updates its current position estimate by adding the displacement inferred from its own motion. Since the update is recursive, even small errors can accumulate over long trajectories, leading to drift when no external sensory correction is available.

At the neural population level, grid-cell activity is often described using continuous attractor models. Under this view, the population state evolves smoothly as the animal moves, and spatial displacement corresponds to a translation of the activity pattern. Because grid-cell firing is periodic, the resulting population representation can be modeled geometrically as a compact periodic manifold, such as a torus. This manifold structure provides a useful constraint for path integration: nearby movements produce nearby changes in neural activity, while the global periodic structure supports stable multi-scale spatial coding. Place-cell activity can complement this metric code by providing context-specific anchoring, helping correct accumulated drift when landmarks or other sensory cues are available.

D LIMITATION

While our brain-inspired framework demonstrates performance gains and provides analytical insights, several aspects warrant further investigation. One consideration is that constructing cognitive maps externally involves a trade-off between memory usage and capacity. Additionally, the cognitive map is currently implemented as an external module rather than fully integrated into the world model's

intrinsic architecture. This decoupled design facilitates analytical clarity and stabilizes the latent space of the frozen CDiT, but future work could explore integrating compositional spatial representations directly within the model to further enhance internal reasoning and generalization.

E Experiment details and Additional Results

E.1 Experimental Hyperparameters settings

In our experimental settings, the **prediction imagination module** employs a frozen NWM CDiT-XL model. For the **compositional cognition module**, we use grid modules with spatial periods λ_i set to $\{3, 4, 5, 7\}$. We initialize 18 random grid-to-place projection matrices W_{gp} to implement the reactivation gating mechanism, with a sparsity mask strength of $z = 0.4$.

E.2 Evaluation Metrics Details

Generation Performance. We evaluate generation performance using PSNR Hore & Ziou (2010), LPIPS Zhang et al. (2018), and DreamSim Fu et al. (2023) for per-sample fidelity, and assess distribution-level visual quality with FID Heusel et al. (2017) and FVD Unterthiner et al. (2019).

Peak Signal-to-Noise Ratio (PSNR) is defined as

$$\mathrm{PSNR}(x, \hat{x}) = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}(x, \hat{x})} \right), \tag{21}$$

where MSE denotes mean square error and L denotes the maximum possible pixel value.

Learned Perceptual Image Patch Similarity (LPIPS) is computed as

$$\mathrm{LPIPS}(x, \hat{x}) = \sum_l \frac{1}{H_l W_l} \left\| w_l \odot (\phi_l(x) - \phi_l(\hat{x})) \right\|_2^2, \tag{22}$$

where H_l and W_l denote the height and width of the feature map at layer l respectively, $\phi_l(\cdot)$ denotes features extracted from layer l of a pretrained network and w_l are learned channel-wise weights.

DreamSim is defined based on semantic similarity as

$$\mathrm{DreamSim}(x, \hat{x}) = 1 - \cos\big(f(x), f(\hat{x})\big), \tag{23}$$

where $f(\cdot)$ denotes the embedding network used by DreamSim.

To evaluate distribution-level visual quality, we employ Fréchet Inception Distance (FID),

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right), \tag{24}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) denote the mean and covariance of real and generated samples in the Inception feature space. For video generation, we further report Fréchet Video Distance (FVD), which adopts the same formulation as FID but computes statistics using features extracted from a pretrained video recognition model (e.g., I3D), capturing both spatial and temporal dynamics.

Navigation Performance. We evaluate navigation performance using Absolute Trajectory Error (ATE) to quantify global trajectory accuracy and Relative Pose Error (RPE) to capture local pose consistency between consecutive steps Sturm et al. (2012).

Absolute Trajectory Error (ATE) measures the global discrepancy between the predicted trajectory and the ground-truth trajectory after alignment, and is defined as

$$\mathrm{ATE} = \frac{1}{T} \sum_{t=1}^{T} \left\| \mathbf{p}_t - \hat{\mathbf{p}}_t \right\|_2, \tag{25}$$

where $\mathbf{p}_t$ and $\hat{\mathbf{p}}_t$ denote the ground-truth and predicted positions at timestep t, respectively, and T is the trajectory length.

Relative Pose Error (RPE) evaluates local pose consistency by measuring the error in relative motion between consecutive poses, defined as

$$\mathrm{RPE} = \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| (\mathbf{p}_{t+1} - \mathbf{p}_t) - (\hat{\mathbf{p}}_{t+1} - \hat{\mathbf{p}}_t) \right\|_2. \tag{26}$$

E.3 Topological Data Analysis Methodology

We use persistent homology to quantify the latent spatial organization of diffusion-based navigation world models. For each CDiT block ℓ, we collect the hidden states generated along a trajectory and form a point cloud

$$H^{(\ell)} = \{h_1^{(\ell)}, \dots, h_N^{(\ell)}\}, \qquad h_i^{(\ell)} \in \mathbb{R}^d.$$

Each point corresponds to one latent state during generation or autoregressive navigation rollout. We then analyze whether these latent states form stable low-dimensional topological structures.

Since the original hidden states are high-dimensional and may contain nuisance appearance or semantic factors, we first reduce them to 16 dimensions using PCA, and then apply Isomap to obtain a 10-dimensional intrinsic embedding:

$$\tilde{H}^{(\ell)} = \mathrm{Isomap}_{10}\big(\mathrm{PCA}_{16}(H^{(\ell)})\big).$$

PCA is used as a denoising step, while Isomap helps preserve intrinsic geodesic structure when the latent states lie on a curved manifold. For visualization only, we project the point cloud to three dimensions; all persistence scores are computed from the 10-dimensional embedding.

Given the reduced point cloud $\tilde{H}^{(\ell)}$, we construct a Vietoris–Rips filtration. For a scale parameter ϵ, the complex is defined as

$$\mathrm{VR}_\epsilon(\tilde{H}^{(\ell)}) = \left\{\sigma \subseteq \tilde{H}^{(\ell)} : \|x_i - x_j\|_2 \le \epsilon,\ \forall x_i, x_j \in \sigma\right\}.$$

As ϵ increases, topological features appear and disappear. For each feature f, its persistence is defined as

$$\mathrm{Pers}(f) = \mathrm{death}(f) - \mathrm{birth}(f).$$

Features with larger persistence are considered more stable across scales and less likely to be caused by local sampling noise.

We focus on the first homology group H_1, which captures loop-like structures. For each CDiT block, we summarize the persistence diagram by the maximum H_1 lifetime:

$$S^{(\ell)} = \max_{f \in H_1} \mathrm{Pers}(f).$$

A larger $S^{(\ell)}$ indicates a stronger loop-like latent structure. Following our main analysis, a block is considered to contain a significant cognitive-map-like representation when

$$S^{(\ell)} > \tau,$$

where $\tau = 0.1$ is used as the default threshold.

The robustness of this pipeline is supported by the stability theorem of persistent homology Chazal et al. (2014):

$$d_b\big(\mathrm{Dgm}(\mathrm{Rips}(X)), \mathrm{Dgm}(\mathrm{Rips}(Y))\big) \le d_{GH}(X, Y),$$

where d_b denotes bottleneck distance and d_{GH} denotes Gromov–Hausdorff distance. This implies that if preprocessing approximately preserves the intrinsic metric structure, the resulting persistence diagrams remain stable. We therefore use PCA mainly to suppress high-dimensional noise and Isomap to preserve intrinsic manifold geometry. Additional robustness checks under different preprocessing choices are reported in the appendix.

We interpret persistent H_1 features as evidence of loop-like latent spatial organization, rather than as a literal biological grid-cell torus. In our goal-directed navigation setting, such S^1-like topology can be viewed as a task-adaptive abstraction of spatial progression along trajectories.

E.4 Additional Ablation Analyses.

1. Do cognitive-map-like representations also emerge in unseen environments? Fig. 10 illustrates how the lifetimes of the most prominent H1 features vary across CDiT blocks in unseen environments. Notably, blocks that exhibited the longest persistence in known environments (around blocks 8–13) show reduced lifetimes in unseen environments, while other intermediate blocks

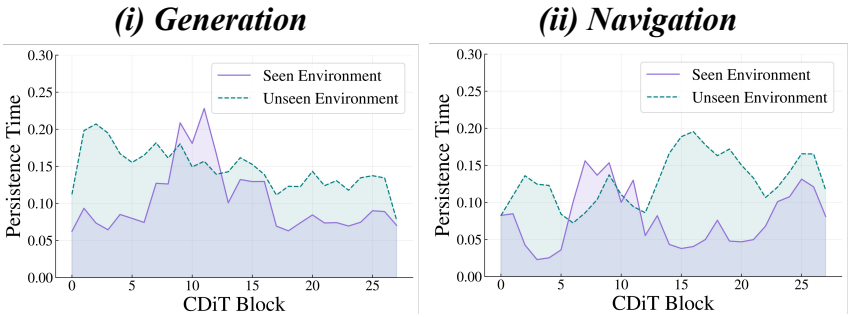

Figure 10: **Persistence in Known vs. Unseen Environments.** The figure compares the persistence of cognitive-map-like representations in known and unseen environments under settings (i) and (ii).

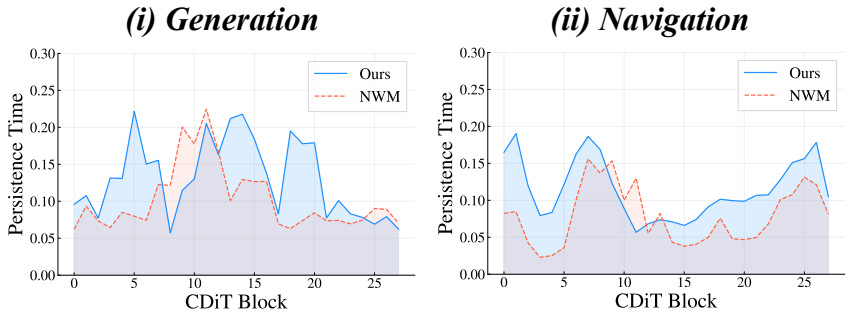

Figure 11: **Impact of the Framework on Representation Persistence.** Persistence of cognitive-map-like representations is compared before and after incorporating our framework under settings (i) and (ii).

(around blocks 16–18) display increased persistence. This indicates that the positions of dominant cognitive-map-like representations shift when the model encounters novel environments.

We interpret this phenomenon as follows. In known environments, the model's representations often entangle environment-specific cues with spatial structure, leading it to memorize more local environmental features rather than capturing the underlying spatial layout. In contrast, unseen environments lack familiar cues, forcing the model to rely on more abstract and invariant spatial representations. This shift enables other blocks to develop robust cognitive-map-like representations, reflecting the model's capacity to generalize spatial structure beyond memorized environmental details. The observed redistribution of persistence highlights the adaptive and task-relevant nature of the latent spatial representations in DWMs.

2. Can our method enhance the formation of intrinsic cognitive-map-like representations? Fig. 11 illustrates how the lifetimes of the most prominent H1 features vary across CDiT blocks before and after applying our method. To achieve smoother results, we applied a moving window. Our approach substantially enhances cognitive-map-like representations persistence in both settings and exhibits superior stability over the original method in single-step generation tasks.

Based on the previous analysis, this also indicates that our method helps the model learn the underlying spatial structure and state transitions, rather than forcing it to memorize more features of known environments.

E.5 Periodic Ablation Analysis of the Compositional Cognition Module.

In the **compositional cognition module**, each grid module has a spatial period λ_i. To achieve exponentially high representational capacity, the periods of different modules must be coprime.

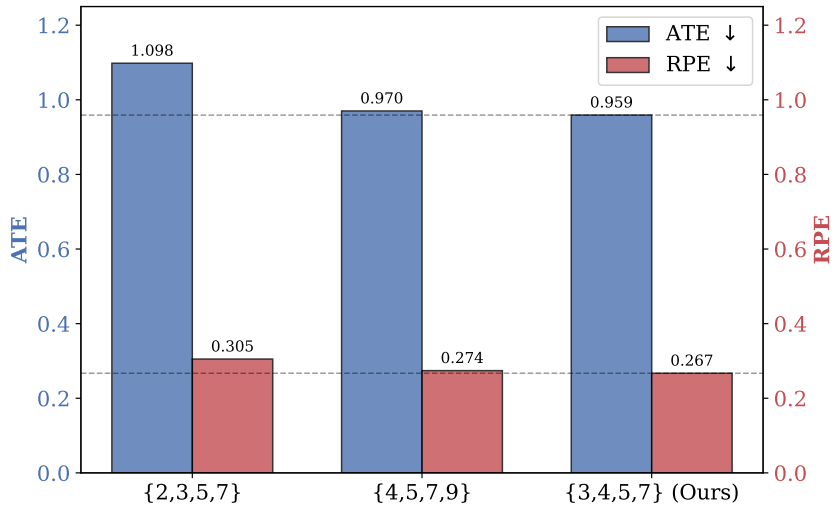

Figure 12: **Navigation performance under different choices of periods.** We report navigation evaluation metrics ATE and RPE under different periods, our choice of periods achieves the best performance.

When the periods are coprime, the combination of modules can generate $\prod_i \lambda_i$ unique states without aliasing. If the periods share common factors, the combined states overlap periodically, reducing the number of distinguishable spatial states and thus limiting the capacity of the cognitive map.

In our experiments, we adopt the coprime period set $\{3, 4, 5, 7\}$ as the primary configuration to maximize the number of distinguishable spatial states while maintaining controlled memory consumption. For comparison, we also evaluate alternative non-optimal sets such as $\{2, 3, 5, 7\}$ and $\{4, 5, 7, 9\}$. The results indicate that the set $\{3, 4, 5, 7\}$ consistently achieves optimal performance in generation and navigation tasks, whereas other sets fail to balance capacity and memory effectively, leading to reduced predictive performance and increased error accumulation.

Our analysis indicates that for the period set $\{3, 4, 5, 7\}$, the required number of grid cells is $N_g = 99$. In comparison, the set $\{4, 5, 7, 9\}$ requires $N_g = 171$ grid cells, leading to approximately a $2\times$ increase in memory consumption. However, this substantial increase in memory does not yield a commensurate improvement in empirical performance. This observation suggests that representational capacity is not solely determined by the raw number of grid cells, but is also constrained by the efficiency with which memory resources are allocated across different spatial scales. Navigation performance comparison in Fig. 12.

For the set $\{2, 3, 5, 7\}$, the number of grid cells is $N_g = 87$, which is of a similar order of magnitude to that of $\{3, 4, 5, 7\}$. Nevertheless, its effective representational capacity is significantly lower (approximately $4\times$), a discrepancy that is consistently reflected in the experimental results. This highlights that an appropriate choice of periods must jointly account for both representational efficiency and memory utilization, rather than optimizing either factor in isolation.

E.6 Cognitive-map-like Representations Under Different Model Settings

We evaluate the emergence of internal representations in the NWM model under four settings: Denoise and Predict, each with linear or squared cosine schedules. Denoise learn how to predict noise, while Predict directly estimates the true latent vector. The linear schedule reduces noise uniformly, whereas the squared cosine schedule decreases it slowly at first and faster later. Experiments focus on intermediate diffusion timesteps.

In our experiments, we performed 60 random samplings, for each trial, we count features exceeding 0.1 and normalize by the 28 CDiT blocks to quantify significance (Fig. 13).

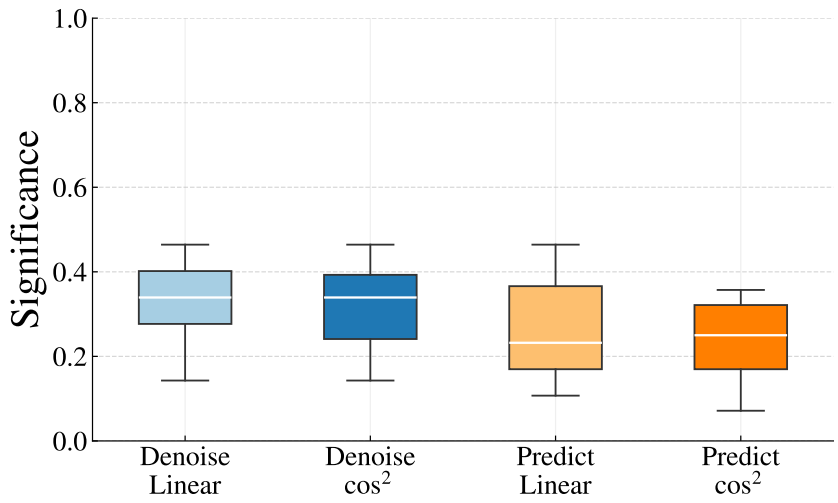

Figure 13: **Cognitive-map-like representations under different settings.** Significance of internal representations in the NWM model across Denoise and Predict objectives with linear or squared cosine schedules.

The choice of schedule has little effect, while Denoise shows higher significance than Predict (≈ 0.1 difference). Denoising emphasizes spatial structure by predicting noise, whereas Predict focuses on original vector details, which may hinder spatial representations. From a neuroscience perspective, denoising aligns more closely with human spatial imagination Park et al. (2024), suggesting that the emergence of such representations may reflect consistency between the model's objective and human cognitive processes.

E.7 Statistical Analysis of Cognitive-map-like Representations

To ensure the stability and reliability of our analysis, we conducted a statistical evaluation of the results. In our experiments, we performed 60 random samplings under both generation and navigation settings to assess the consistency and robustness of representations across trials.

Representation Stability Analysis. We first analyzed the stability of map-like representations in the original diffusion-based navigation world model. For each trial, we counted the number of features exceeding a threshold of 0.1 and normalized it by the total number of CDiT blocks (28) to quantify significance. We then compared the stability before and after incorporating our method. The results are shown in Fig. 14. In both settings of NWM, the mean significance remained below 0.4, indicating unstable representations. Incorporating our method substantially increased both the emergence and stability of these representations.

Unseen Environment Representation Analysis. Next, we analyzed differences in map-like representations of the diffusion-based navigation world model between known and unseen environments. We measured the differences in persistence at corresponding layers, as shown in Fig. 15. We observe that representations are more pronounced in the known environment around layers 8–13, whereas the unseen environment exhibits stronger representations in the early and late layers. We explain this by noting that, in known environments, learned representations may entangle environment-specific visual or statistical cues with spatial structure, leading the model to memorize more environmental features rather than capturing the underlying spatial structure. In contrast, unseen environments force the model to rely on more abstract and invariant spatial representations, which enables the emergence of more robust cognitive-map-like representations.

Effect of Our Method on Representational Stability. Finally, we analyzed the duration of the most persistent cognitive-map-like representations across model layers after incorporating our method.

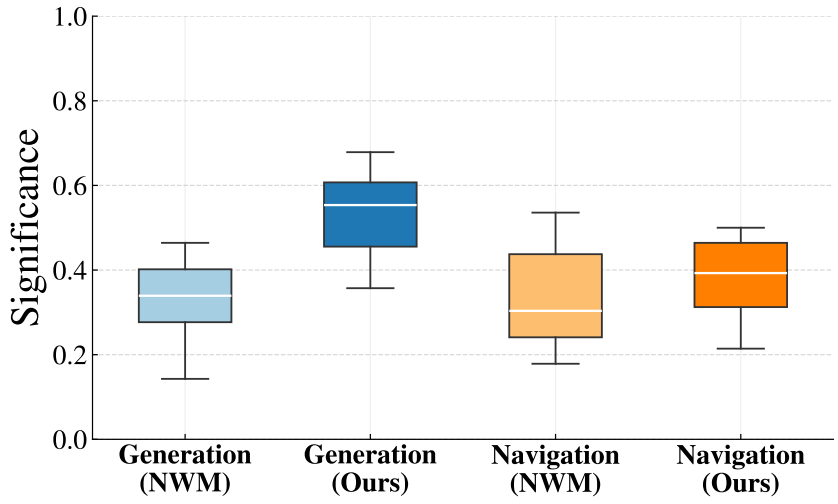

Figure 14: **Stability analysis of cognitive-map-like representations**. The original NWM shows low-persistence, unstable features (mean < 0.4), while our method significantly improves their emergence and robustness.

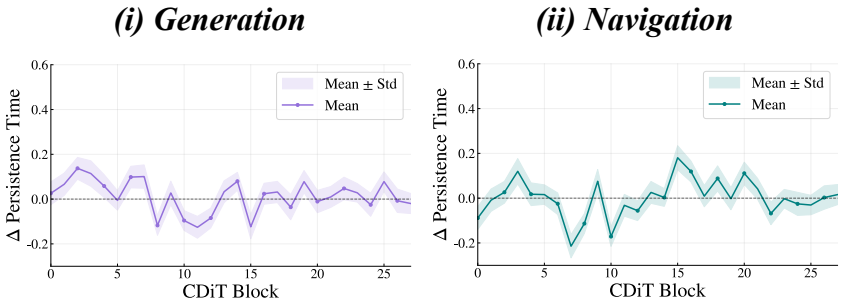

Figure 15: **Persistence differences of cognitive-map-like representations across known and unseen environments.** Known environment shows stronger features around mid layers, while unseen environment exhibits stronger features in early and late layers.

In the experiments, we measured the differences in persistence time across layers as the metric for visualization across different settings and environments. The results, shown in Fig. 16, indicate that our method consistently increases the robustness and emergence of these representations.

E.8 Additional Experiment Results

For generation tasks, we report FVD on additional datasets to evaluate video generation performance (results in Table 5), while for navigation tasks, we report ATE and RPE as metrics of trajectory accuracy and pose consistency (results in Table 6). We also present the performance of our method on all datasets, as shown in the following figures.

E.9 Additional Manifold Visualization

To further examine the internal representations, Fig. 22 visualizes additional cognitive-map-like patterns that emerge in both the NWM and our method. These visualizations demonstrate that our method not only stabilizes previously observed map-like representations but also enhances the

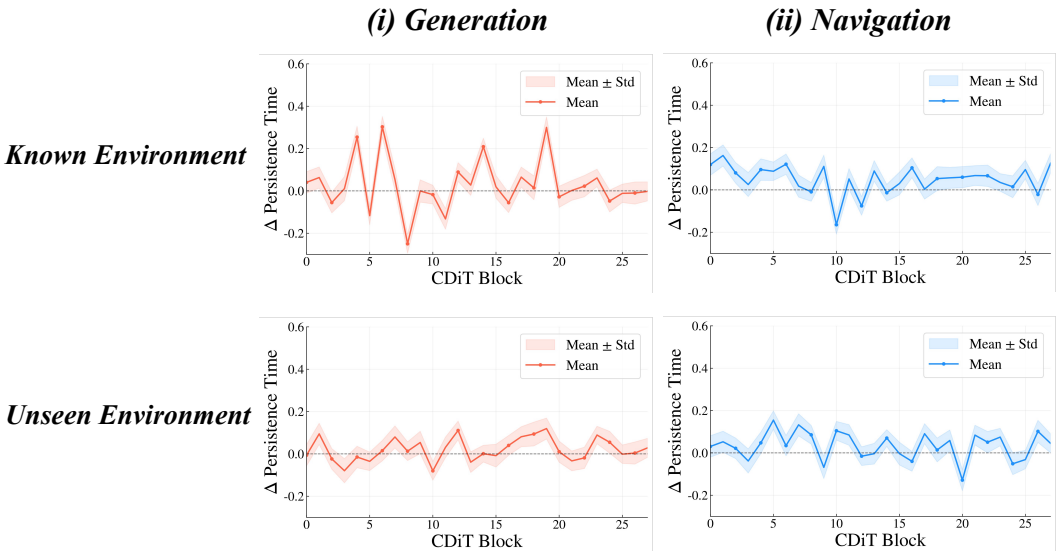

Figure 16: **Persistence differences of cognitive-map-like representations across layers.** Differences in persistence time across layers indicate increased robustness and emergence of these representations with our method.

Table 5: **Generation Performance Comparison.** 16s video generation performance of different models across multiple datasets (FVD).

Model	RECON	TartanDrive	SCAND	HuRoN
DIAMOND	725.549	2351.726	1921.385	862.971
NWM	191.096	510.199	387.696	268.297
Ours	**169.855**	**495.231**	**371.209**	**251.125**

Table 6: **Navigation Performance Comparison.** Navigation performance comparison of different models across multiple datasets (ATE and RPE).

Model	RECON		TartanDrive		SCAND		HuRoN	
	ATE ↓	RPE ↓	ATE ↓	RPE ↓	ATE ↓	RPE ↓	ATE ↓	RPE ↓
GNM	1.912	0.723	6.713	1.598	2.187	0.625	**3.625**	1.024
NoMaD	1.875	0.509	6.401	1.302	2.225	0.512	3.717	0.996
NWM	1.057	0.300	5.729	1.209	1.234	0.313	3.931	0.968
Ours	**0.959**	**0.267**	**5.657**	**1.157**	**1.188**	**0.294**	3.729	**0.951**

emergence of new, persistent spatial features, supporting improved long-horizon prediction and navigation performance.



Figure 17: **RECON Performance.** Performance of our method on the RECON dataset.



Figure 18: **Tartan Drive Performance.** Performance of our method on the Tartan Drive dataset.

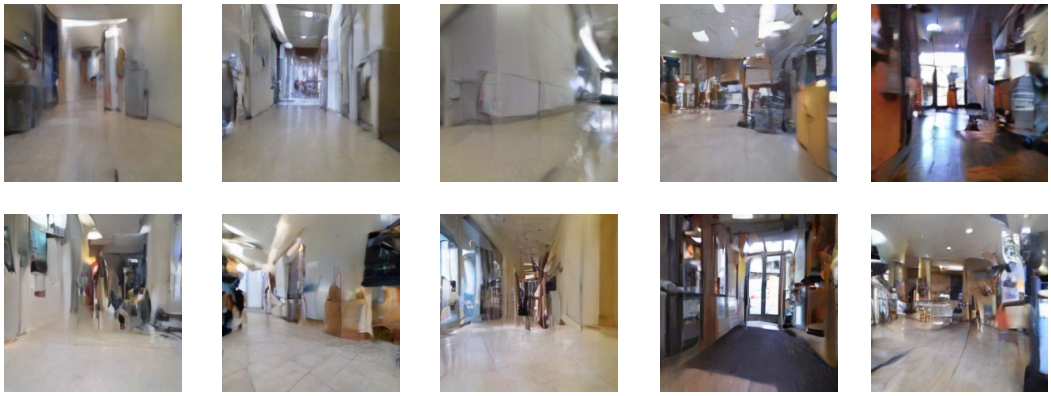

Figure 19: **SCAND Performance.** Performance of our method on the SCAND dataset.



Figure 20: **HuRoN Performance.** Performance of our method on the HuRoN dataset.

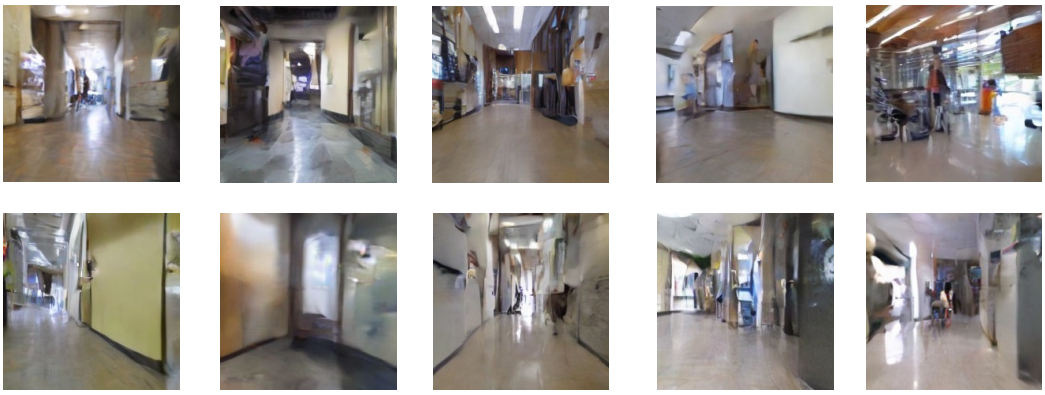

Figure 21: **Go Stanford Performance.** Performance of our method on the Go Stanford dataset.

Generation (NWM)

Navigation (NWM)

Generation (Ours)

Navigation (Ours)

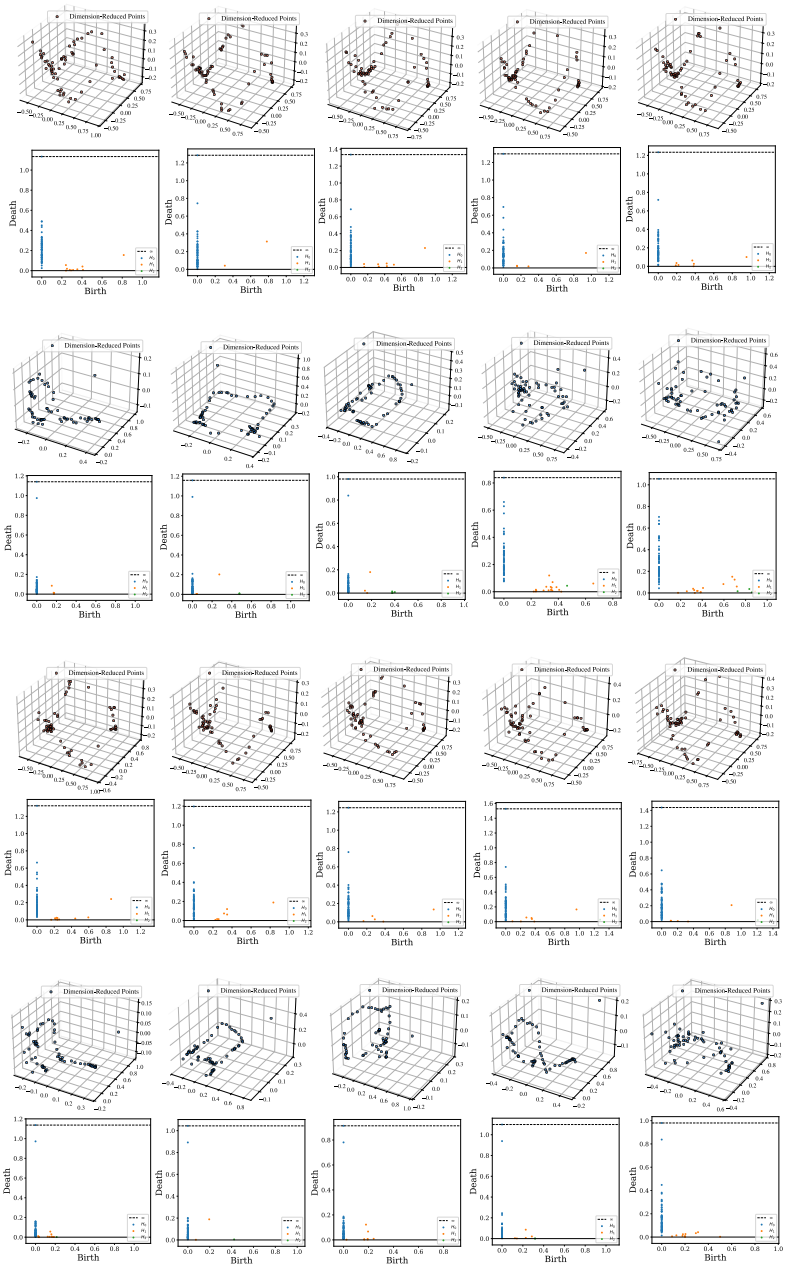

Figure 22: **Representation Visualization.** Visualization of additional cognitive-map-like representations emerging in the NWM and Ours.