RAE-NWM: Navigation World Model in Dense Visual Representation Space

Mingkun Zhang[1], Wangtian Shen[1], Fan Zhang[2], Haijian Qin[3], Zihao Pei[1], and Ziyang Meng[1⋆]

[1] Department of Precision Instrument, Tsinghua University, Beijing, China
zhangmk24@mails.tsinghua.edu.cn, ziyangmeng@tsinghua.edu.cn
[2] University of Rochester, Rochester, NY, USA
[3] Beijing Information Science and Technology University, Beijing, China

Abstract. Visual navigation requires agents to reach goals in complex environments through perception and planning. World models address this task by simulating action-conditioned state transitions to predict future observations. Current navigation world models typically learn state evolution under actions within the compressed latent space of a Variational Autoencoder, where spatial compression often discards fine-grained structural information and makes precise action-conditioned prediction more difficult. To better understand the propagation characteristics of different representations, we conduct a linear dynamics probe and observe that dense DINOv2 features exhibit stronger linear predictability for action-conditioned transitions. Motivated by this observation, we propose the Representation Autoencoder-based Navigation World Model (RAE-NWM), which generatively models navigation dynamics in a dense visual representation space. We employ a Conditional Diffusion Transformer with a Decoupled Diffusion Transformer head (CDiT-DH) to model continuous transitions, and introduce a separate time-driven gating module for dynamics conditioning to regulate action injection strength during generation. Extensive evaluations show that modeling sequential rollouts in this space improves structural stability and action accuracy, benefiting downstream planning and navigation. Code is available at https://github.com/20robo/raenwm.

Keywords: World Models · Visual Navigation · Flow Matching · Visual Representation Learning

1 Introduction

Visual navigation is a fundamental problem in autonomous robotics, requiring agents to perceive their surroundings and reach target goals safely in complex environments [10, 55]. End-to-end learning-based solution [32, 38] has emerged as an effective approach for this task, but it often struggles with uninterpretable intermediate decision processes and is difficult to incorporate additional constraints after training. Navigation World Models (NWM) [5], on the other hand,

⋆ Corresponding author

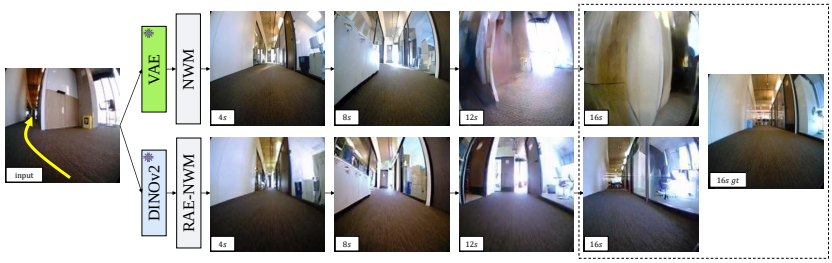


Fig. 1: Long-horizon generation comparison. We compare sequential rollouts from the baseline VAE-based NWM versus our DINO-based RAE-NWM. The VAE-based model exhibits severe structural degradation at later horizons ($k = 12s, 16s$), whereas RAE-NWM maintains strong structural integrity throughout the sequence.

offer an explicit solution with a flexible structure to address these issues by evaluating candidate trajectories through environment simulation. Using generative models conditioned on historical states and physical actions, NWM predicts future observations to assess navigation safety and goal progress. In the context of visual navigation, since these predictions directly inform how to plan, the world model must do more than simply synthesize high-fidelity images. It is crucial that the model maintains consistent spatial geometric stability and precise action controllability over extended horizons; otherwise, the reliability of the resulting decisions drops significantly. However, most existing methods [36, 51] perform these predictive rollouts within the latent space of a Variational Autoencoder (VAE) [22]. While such a spatial compression represents an inherent mechanism of this encoding process, it typically limits the preservation of critical geometric information from the original observations [29, 53]. Consequently, during long-horizon future predictions, this lack of structural consistency often causes VAE-based methods to experience significant structural collapse and kinematic deviation, limiting their reliability for downstream path planning (Figure 1).

To address the spatial degradation introduced by VAE-based latent spaces, it is important to consider representation spaces that preserve geometric structure. DINOv2 [26] contains rich spatial semantics and geometric structures, as evidenced by its strong performance in various recent vision tasks [24, 43]. Some recent studies [3, 52, 54] attempt to use self-supervised or representative visual features for environment simulation, showing the promise of feature-space world modeling. However, existing feature-space predictive models are mainly explored through deterministic feature prediction, often in manipulation or general video prediction/planning settings, rather than generative, action-conditioned rollouts for first-person visual navigation. Continuous generative models offer a suitable framework for modeling these smooth temporal transitions. Given the inherent cost of training such heavily parameterized networks, it is prudent to first verify how well the semantic representation space supports action-conditioned dynamics. Therefore, we introduce a Linear Dynamics Probe to investigate this

property. Our analysis shows that DINOv2 features exhibit strong linear predictability for action-conditioned state transitions. Inspired by this observation, we build our generative transition model entirely in this dense visual representation space.

In particular, we propose the Representation Autoencoder-based Navigation World Model (RAE-NWM) in this paper. Following the RAE [53] approach, our model extracts uncompressed visual tokens using a frozen DINOv2 encoder and reconstructs the final observations using a frozen pretrained RAE decoder. The decoder is used only as a fixed reconstruction component, and only the transition model between the frozen encoder and decoder is trained. Specifically, we train a Conditional Diffusion Transformer with a Decoupled Diffusion Transformer head (CDiT-DH) to serve as the generative backbone [27, 44]. Motivated by the coarse-to-fine nature of diffusion-style generative processes [4, 16], we introduce a dynamics conditioning module equipped with a time-driven dynamic gating mechanism. Instead of standard additive conditioning, this module adaptively adjusts the strength of kinematic conditioning throughout the probability flow. This mechanism balances high-fidelity visual generation and precise action controllability, preserving global geometric topology while refining local structural details.

Our contributions can be summarized as follows. First, we study generative navigation world modeling in dense visual representation spaces instead of compressed VAE-based latent spaces. This formulation preserves richer spatial structure and provides a suitable representation space for modeling action-conditioned first-person navigation dynamics. Second, we develop a generative transition architecture for navigation world models built on CDiT-DH and an adaptive gating mechanism. This design enables stable modeling of high-dimensional visual representations while maintaining global geometric consistency and local structural details. Third, extensive experiments verify that our approach improves long-horizon rollout stability and achieves stronger performance in both open-loop trajectory evaluation and downstream navigation planning tasks.

2 Related Work

Visual Representation Space. Traditional generative paradigms typically construct their representation space using compression models. Standard approaches, from the foundational Variational Autoencoder (VAE) [22, 29] to methods incorporating semantic alignments or masked reconstruction objectives [7, 8, 46, 47], introduce a spatial compression bottleneck. Relying on these low-dimensional latents limits low-level geometric fidelity, even with semantic enhancements. To address this limitation, Representation Autoencoders (RAE) [53] reconstruct images directly from uncompressed visual foundation models [14, 26, 42]. Although such representations are often considered to primarily encode high-level semantic information rather than low-level visual details [41, 48], RAE shows that they can nevertheless support high-fidelity spatial reconstruction when paired with a vision transformer (ViT) [9] decoder, without relying on compression.

World Models for Embodied Intelligence. Unlike end-to-end visual navigation policies [28,31–33,35,38], world models [12,23] explicitly simulate environmental dynamics for decision-making. Recent Navigation World Models [5,36] condition future predictions on physical actions. However, these frameworks often rely on VAE-based latent spaces, which can degrade spatial structure. To mitigate this issue, some methods [49, 51] explicitly project past frames into future views, but this increases computational complexity. Other works explore hybrid architectures that combine autoregressive generation with diffusion chunking [34] to extend rollout horizons, or study representation-space world models based on self-supervised or representative visual features [3,52,54]. DINO-WM [54] learns dynamics in continuous DINOv2 feature space through deterministic feature prediction, while V-JEPA 2 [3] studies self-supervised video prediction and planning. Cloning Deterministic Worlds [45] further discusses limitations of pixel-reconstructing VAE representations for navigation world modeling. These works demonstrate the potential of representation-level simulation, but they mostly rely on deterministic prediction or settings such as manipulation and general video prediction, rather than generative first-person navigation rollouts. Since navigation dynamics follow smooth action-induced geometric changes, RAE-NWM models them with continuous-time generative dynamics in dense representation space for stable long-horizon planning.

3 Task Definition and Representation Analysis

3.1 Task Definition

A world model couples an encoder, which maps complex raw observations into a tractable state space, with a predictor that forecasts future states based on agent actions. In planar visual navigation tasks, an agent receives a first-person RGB observation $\mathbf{o}_i \in \mathbb{R}^{H\times W\times 3}$ at step i. We map a sequence of m context frames $\mathbf{o}_{\text{cond},i} = \mathbf{o}_{i-m+1:i}$ through an encoder to obtain the state representation $\mathbf{z}_{\text{cond},i} = \text{Enc}(\mathbf{o}_{\text{cond},i})$, capturing the current environmental context. Following NWM [5], we introduce a relative temporal shift k to specify the prediction horizon. We then represent the agent motion over this interval using a planar motion condition $\boldsymbol{a}_{i\to i+k} = [u_x, u_y, \omega]_{i\to i+k}$, where u_x, u_y and ω denote the planar translation and yaw rotation respectively. Together, k and $\boldsymbol{a}_{i\to i+k}$ describe the kinematic condition of the agent during this period. The objective of the world model for visual navigation tasks is to learn a transition dynamics function within this representation space:

$$\hat{\mathbf{z}}_{i+k} = \text{Pred}_\theta(\mathbf{z}_{\text{cond},i}, \boldsymbol{a}_{i\to i+k}, k) \tag{1}$$

3.2 Representation Analysis

To effectively learn the transition function defined in Equation 1, it is critical to determine an appropriate state representation space to encode raw observations. This choice determines whether the subsequent predictor can readily

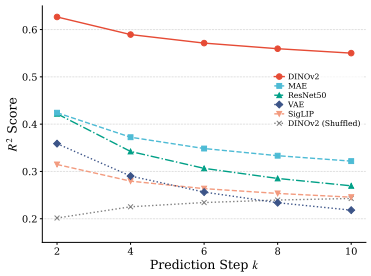


Fig. 2: Linear predictability of action-conditioned state transitions across various visual representation spaces, measured by the global R^2 score over the prediction horizon step k.

extract action-conditioned dynamics from the representation, or instead has to model complex visual variations that are unrelated to the underlying motion. Inspired by linear probing techniques [1], we introduce a linear dynamics probe to evaluate the kinematic predictability within various representation spaces. To evaluate such predictability without involving generative modeling, we freeze the pre-trained encoder and only train a linear probe in its representations. We extract the spatial token representation $\mathbf{z}_i$ from the current observation and use the future representation $\mathbf{z}_{i+k}$ as the prediction target. The probe predicts the future state conditioned on the intermediate motion $\boldsymbol{a}_{i\to i+k}$ through a linear formulation:

$$\hat{\mathbf{z}}_{i+k} = \mathbf{z}_i + \boldsymbol{a}(\mathbf{z}_i) + \mathbf{B}(\boldsymbol{a}_{i\to i+k}) \tag{2}$$

where $\boldsymbol{a}$ and $\mathbf{B}$ denote learnable linear transformations. We restrict the probe to this simple setup to ensure the results directly reveal whether the underlying feature space natively supports action-conditioned dynamics.

The linear probes are trained and evaluated on the navigation datasets detailed in Section 5.1. To quantify this linear predictability, we report the global R^2 score [13], which measures how well future states can be explained by a linear function of the current state and action. As indicated by Figure 2, the uncompressed token space of DINOv2 maintains consistently high predictability across the entire prediction horizon. In contrast, VAE-based compressed representations yield substantially lower scores. Other common visual encoders, such as those optimized for pixel-level masked reconstruction (e.g., MAE [14]) or global semantic alignment (e.g., SigLIP [42] and ResNet50 [15]), also perform poorly. Furthermore, to study whether the high scores of DINO tokens arise mainly from its representation capability of global semantic information, we introduce a spatially shuffled DINOv2 baseline, where the spatial order of the encoded tokens is randomly permuted before being fed to the probe. This operation disrupts the spatial structure of the representation, and the resulting performance drops substantially compared to the original DINO features. This observation suggests that the performance gap comes from the spatial structure of DINOv2 features,

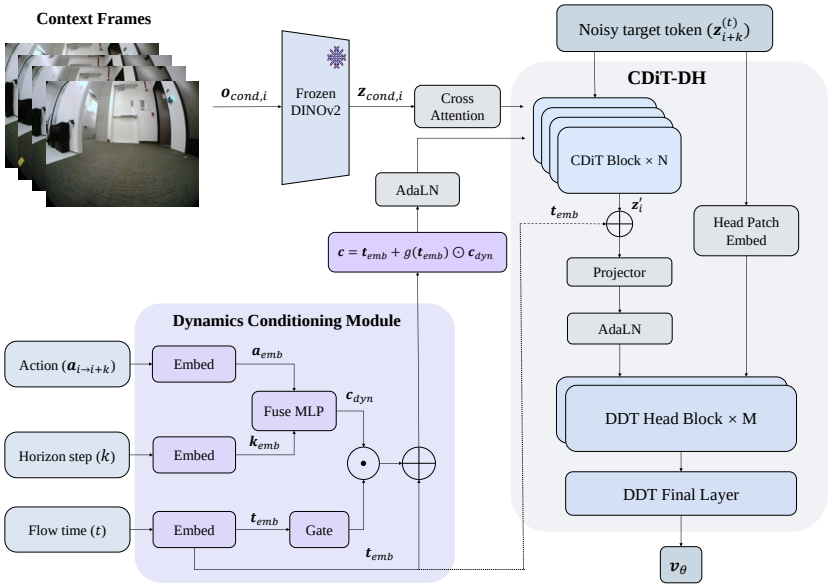


Fig. 3: Architecture of the proposed RAE-NWM. Our model encodes context frames into a sequence of tokens $\mathbf{z}_{\text{cond},i}$ via a frozen DINOv2. The dynamics conditioning module then integrates the agent motion $\boldsymbol{a}_{i\to i+k}$ and prediction horizon k with the flow time t through a time-driven gating mechanism. Finally, the CDiT-DH backbone predicts the flow velocity $\mathbf{v}_\theta$.

rather than static global semantics alone. Therefore, we build the world model directly in this dense representation space.

4 RAE-NWM

We formulate the proposed Representation Autoencoder-based Navigation World Model (RAE-NWM) to parameterize the transition dynamics function defined in Section 3. This section details the network architecture and the training pipeline. We first specify the visual state representation extraction (Sec. 4.1) and introduce the generative backbone CDiT-DH (Sec. 4.2). We then describe the time-driven dynamics conditioning module designed for adaptive kinematic modulation (Sec. 4.3), and finally describe the training objective and sequential rollout procedure (Sec. 4.4).

4.1 State Representation

Given a raw visual observation $\mathbf{o}_i$, we construct its state representation strictly within the dense representation space. We employ a frozen DINOv2 encoder and discard the `[CLS]` token to retain only the uncompressed spatial patch tokens

$\mathbf{z}_i = \mathrm{Enc}(\mathbf{o}_i) \in \mathbb{R}^{L \times d}$. In our implementation, we extract $L = 16 \times 16 = 256$ patch tokens with a feature dimension of $d = 768$. To incorporate temporal context, the features from the context frames are organized into the sequence $\mathbf{z}_{\mathrm{cond},i} = \mathbf{z}_{i-m+1:i}$. This sequence serves as the spatial conditioning input for the generative backbone.

4.2 Generative Backbone: CDiT-DH

As illustrated in Figure 3, after extracting the context representation $\mathbf{z}_{\mathrm{cond},i}$, we employ the CDiT-DH generative backbone. Following the flow matching formulation [25], the CDiT-DH takes as input a noisy target token $\mathbf{z}_{i+k}^{(t)}$. Here $t \in [0, 1]$ denotes the continuous flow time, and the network predicts the corresponding velocity field $\mathbf{v}_\theta$. This prediction is conditioned on the context tokens $\mathbf{z}_{\mathrm{cond},i}$ and a unified global condition vector $\mathbf{c}$ for action and horizon-step control (detailed in Section 4.3).

Deep Generative Backbone. Following the design of NWM [5], we use a deep Conditional Diffusion Transformer (CDiT) composed of N stacked transformer blocks. Given the noisy target token $\mathbf{z}_{i+k}^{(t)}$, the backbone embeds it into patch tokens and adds a 2D sine-cosine positional embedding [14] to provide an explicit spatial inductive bias. Each block applies self-attention to model token-wise spatial dependencies and cross-attention to incorporate the context frames, attending to $\mathbf{z}_{\mathrm{cond},i}$ as keys and values. The global condition $\mathbf{c}$ is injected via Adaptive Layer Normalization (AdaLN) [27] to modulate features throughout the transformer blocks. The generative backbone produces a deep contextual feature $\mathbf{z}'_i$:

$$\mathbf{z}'_i = \mathrm{CDiT}_\theta\Big(\mathbf{z}_{i+k}^{(t)} \mid \mathbf{z}_{\mathrm{cond},i},\, \mathbf{c}\Big). \tag{3}$$

Decoupled Diffusion Transformer (DDT) Head. Modeling target representations in a high-dimensional token space can be challenging for standard diffusion transformer backbones. Following the design principle of RAE [53], we adopt a shallow-and-wide DDT [44] head to predict the final velocity field. This lightweight head allows the model to better handle high-dimensional token representations without significantly increasing computational cost. In practice, the re-embedded noisy input is modulated by spatial AdaLN under the guidance of the deep features produced by the backbone. The predicted velocity field is then computed as

$$\mathbf{v}_\theta = \mathrm{DDT}_\theta\big(\mathbf{z}_{i+k}^{(t)} \mid \mathbf{z}'_i, t\big). \tag{4}$$

4.3 Dynamics Conditioning Module

To condition the generative backbone while preserving the temporal structure of the generative process, we introduce a dynamics conditioning module that adaptively injects the kinematic control signal. We apply Gaussian Fourier embedding [37] to map the input action $\boldsymbol{a}_{i \to i+k}$, the horizon step k, and the continuous

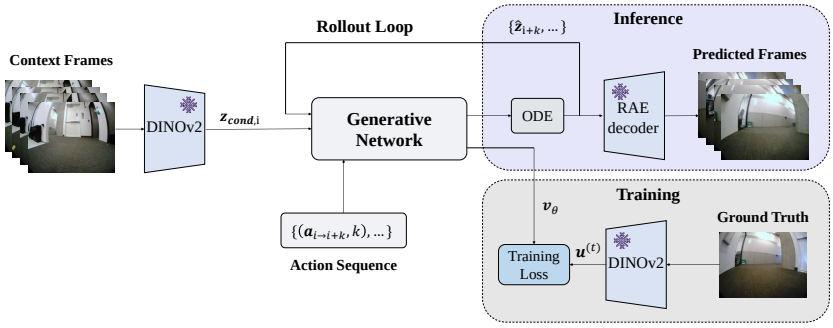


Fig. 4: Training and sequential rollout pipelines of RAE-NWM. During training, the generative network is optimized via a flow matching objective to predict the velocity field $\mathbf{v}_\theta$ matching the target velocity $\mathbf{u}^{(t)}$. During inference, an ordinary differential equation (ODE) solver sequentially generates future states $\{\hat{\mathbf{z}}_{i+k}, \dots\}$ along a given action sequence $\{(\boldsymbol{a}_{i\to i+k}, k), \dots\}$ within a closed representation-space rollout loop. The frozen RAE decoder is applied exclusively for qualitative visualization and pixel-level metric evaluation.

flow time t into a high-frequency latent space. This produces the embeddings $\boldsymbol{a}_{\text{emb}}$, $\mathbf{k}_{\text{emb}}$, and $\mathbf{t}_{\text{emb}}$. We then concatenate $\boldsymbol{a}_{\text{emb}}$ and $\mathbf{k}_{\text{emb}}$ and process them through a multi-layer perceptron (MLP) to extract the dynamics feature $\mathbf{c}_{\text{dyn}}$, which captures the intended motion.

To better balance precise action control and fine-grained visual synthesis within the high-dimensional DINOv2 space, we replace standard additive injection with a time-driven gating function $g(\mathbf{t}_{\text{emb}})$ to dynamically modulate the conditioning strength. This adaptive design aligns with the generation process. Early high-noise stages need strong kinematic priors to establish the global topology, while late low-noise stages require relaxed constraints to refine high-frequency visual details without introducing artifacts. The function $g(\cdot)$ consists of a linear layer and a SiLU activation, followed by a Sigmoid function to constrain the outputs to $(0, 1)$. The final global condition vector $\mathbf{c}$ is formulated as:

$$\mathbf{c} = \mathbf{t}_{\text{emb}} + g(\mathbf{t}_{\text{emb}}) \odot \mathbf{c}_{\text{dyn}} \tag{5}$$

where $\odot$ denotes element-wise multiplication. This modulated condition $\mathbf{c}$ is then injected into the network via AdaLN to guide spatial generation.

4.4 Training Objective and Sequential Rollout

Figure 4 summarizes the training and sequential rollout pipelines of RAE-NWM. This section details the Flow Matching [25] training objective and the representation-space rollout procedure for long-horizon prediction. For clarity, we denote the target representation at a specific horizon as $\mathbf{z}^{(t)}$ and the historical context as $\mathbf{z}_{\text{cond}}$, where t is the flow time. During training, we construct a continuous probability path that transports the clean representation $\mathbf{z}^{(0)}$ to standard

Gaussian noise $\mathbf{z}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We define a linear interpolation path for any $t \in [0, 1]$ as $\mathbf{z}^{(t)} = (1-t)\mathbf{z}^{(0)} + t\mathbf{z}^{(1)}$, which yields the target velocity field $\mathbf{u}^{(t)} = \frac{d\mathbf{z}^{(t)}}{dt} = \mathbf{z}^{(1)} - \mathbf{z}^{(0)}$. The generative network is optimized by matching the predicted velocity $\mathbf{v}_\theta$ directly to this target:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, t}\left[\left\|\mathbf{v}_\theta(\mathbf{z}^{(t)}, t, \mathbf{z}_{\text{cond}}, \mathbf{c}) - \mathbf{u}^{(t)}\right\|_2^2\right]. \tag{6}$$

During inference for the visual navigation task, long-horizon prediction is executed sequentially within the dense representation space. After encoding the initial context frames to obtain the context tokens $\mathbf{z}_{\text{cond},i}$, we initialize the target token for each prediction interval with standard Gaussian noise $\mathbf{z}^{(1)}$. An ordinary differential equation (ODE) solver computes the probability flow backward from $t = 1$ to $t = 0$, guided by the context and the dynamics condition $\mathbf{c}$, producing the predicted clean token $\hat{\mathbf{z}}_{i+k}$. To predict further into the future along a given action sequence, this newly generated state is appended to the sliding context window to condition the subsequent rollout step. The frozen pre-trained RAE decoder [53] reconstructs these representations into pixel space exclusively for qualitative visualization and pixel-level metric evaluation. For trajectory scoring and planning, we operate directly in the representation space instead of decoding predictions to pixels, avoiding geometric distortions and information loss from pixel reconstruction.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our approach on three complementary real-world robot navigation datasets: SACSoN/HuRoN [19], RECON [30], and SCAND [20]. SACSoN captures dynamic indoor human-robot interactions, RECON provides unstructured open-world and off-road terrains, and SCAND contains demonstrations that follow human social norms. We train a single model on the union of their training splits and report results on held-out trajectories from each dataset. We also collect 1,000 trajectories in Matterport3D [6] using the Habitat simulator [40]. Models for the Habitat experiments are trained exclusively on this dataset and evaluated on unseen trajectories. All datasets are partitioned at the trajectory level to prevent temporal leakage.

Evaluation Metrics. To assess open-loop generation quality, we report the Fréchet Inception Distance (FID) [17] for frame-level realism and measure perceptual similarity using LPIPS [50] and DreamSim [11]. We also introduce a patch-wise DINO feature distance, computed as the mean cosine distance between normalized patch tokens extracted by a frozen DINOv2, to evaluate dense semantics and geometric consistency. To measure trajectory and planning accuracy, we test the utility of predicted observations in a downstream planning task based on Cross-Entropy Method, quantifying trajectory fidelity with Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) [39]. Finally, to evaluate

Table 1: Direct long-horizon prediction quality. Performance comparison between RAE-NWM and NWM on the SACSoN dataset at 4-second and 16-second future horizons, generated directly in a single step bypassing intermediate frames.

Model	LPIPS ↓		DreamSim ↓		DINO Distance ↓		FID ↓	
	4s	16s	4s	16s	4s	16s	4s	16s
NWM	0.407	0.470	0.229	0.281	0.402	0.460	26.15	33.06
RAE-NWM	**0.303**	**0.349**	**0.145**	**0.171**	**0.327**	**0.367**	**15.09**	**15.90**

closed-loop simulation performance in Habitat, we deploy the model as a robotic navigation system and report Success Rate (SR) and Success weighted by Path Length (SPL) [2].

Baselines. We compare our approach against several representative baselines. NWM [5] is a world-modeling framework for real-world visual navigation and serves as our primary baseline for open-loop generation and downstream planning. GNM [32] and NoMaD [38] are end-to-end visual navigation policies trained on multiple datasets. GNM is a goal-conditioned policy, while NoMaD introduces a diffusion policy to unify exploration and navigation. For the simulation-oriented Habitat evaluation, we compare against OmniVLA [18], which builds upon the OpenVLA [21] backbone, and the One-Step World Model (One-Step WM) [36], which uses a 3D U-Net dynamics model for one-step generation.

Implementation Details. Detailed network architecture hyperparameters, training configurations, and computational resources are provided in Section A of the Supplementary Material.

5.2 Open-Loop Generation

To identify long-horizon action controllability from the error accumulation inherent in sequential rollout, we first perform a direct prediction experiment. Instead of iteratively generating intermediate frames, we predict specific future observations directly at 4-second and 16-second horizons based on aggregated action sequences. Table 1 demonstrates that NWM degrades significantly across perceptual and semantic metrics at the 16-second mark. As qualitatively shown in Figure 5a, NWM produces observations that deviate substantially from the ground truth, whereas RAE-NWM preserves much higher geometric fidelity.

Building upon this direct controllability, we further evaluate open-loop generation quality using sequential rollout predictions. Given initial observations and ground-truth action sequences, the model iteratively generates 16-second future trajectories at 4 FPS on unseen test data. Figure 6 shows the temporal evaluation on the SACSoN dataset, alongside the qualitative example provided earlier in Figure 1. Figure 5b expands this qualitative evaluation to sequential rollouts on the RECON and SCAND datasets. Both the continuous rollout curves and

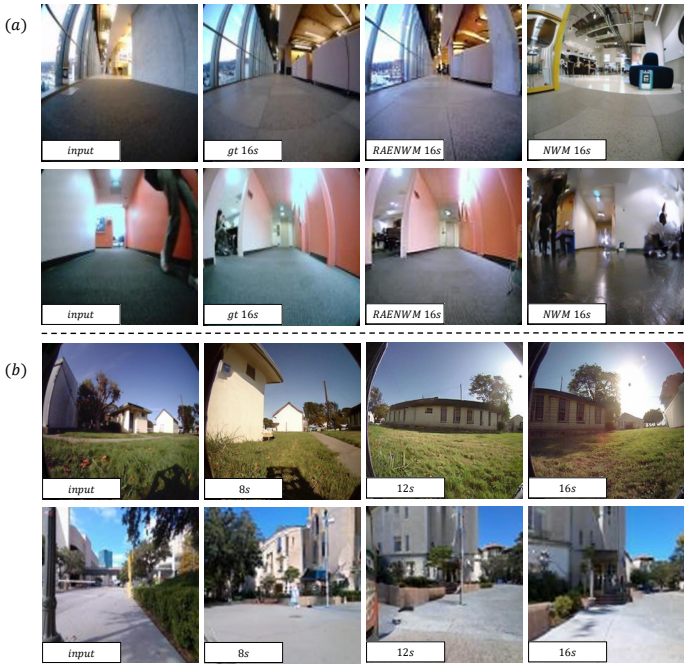
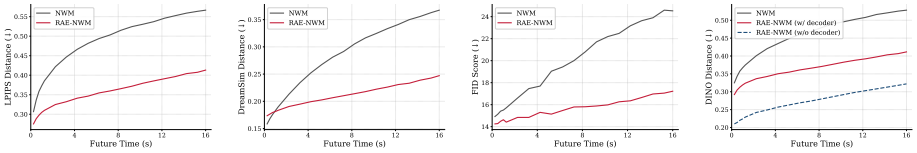


Fig. 5: Qualitative results of long-horizon generation. (a) Direct prediction at the 16-second horizon. NWM produces observations that completely deviate from the ground truth, whereas RAE-NWM maintains high geometric fidelity, demonstrating superior action-conditioned dynamics. (b) Sequential rollouts of RAE-NWM on the RECON and SCAND datasets, exhibiting strong structural consistency and spatial stability over extended horizons.

these visualizations demonstrate that RAE-NWM maintains strong temporal stability and structural consistency over extended horizons. We further inspect rollouts beyond the 16-second horizon and observe smooth LPIPS increases from 16s to 32s on SACSoN, RECON, and SCAND: 0.387→0.438, 0.458→0.517, and 0.523→0.614, respectively, indicating gradual degradation rather than abrupt collapse. The larger increase on SCAND suggests that moving-agent scenes remain more challenging for long-horizon rollout. To rule out the possibility that this performance is an artifact of the pre-trained RAE decoder, we compute the DINO patch distance directly in the token space against the ground truth (Figure 6d). The lower error in this uncompressed representation space confirms the capacity of the model to produce accurate structural predictions.

5.3 Trajectory and Planning Accuracy

To evaluate the standalone goal-conditioned planning performance of RAE-NWM, we use the Cross-Entropy Method (CEM) for trajectory optimization. During each CEM iteration, we first sample 120 candidate action sequences. For



(a) LPIPS **(b)** DreamSim **(c)** FID **(d)** DINO Distance

Fig. 6: Long-horizon sequential rollout quality. Performance comparison between RAE-NWM and the baseline Navigation World Model on the SACSoN dataset over a 16-second future horizon, generated iteratively step-by-step.

Table 2: Trajectory prediction performance and deterministic DINO-token regression ablation. Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) results are reported across all in-domain datasets for trajectories predicted up to 2 seconds. DINO-Reg denotes deterministic DINO-token regression under the same navigation protocol.

Model	SACSoN		RECON		SCAND	
	ATE ↓	RPE ↓	ATE ↓	RPE ↓	ATE ↓	RPE ↓
GNM [32]	3.71	1.00	1.87	0.73	2.12	0.61
NoMaD [38]	3.73	0.96	1.95	0.53	2.24	0.49
NWM [5]	4.12	0.96	**1.13**	**0.35**	1.28	0.33
DINO-Reg	3.14	0.75	1.46	0.36	1.44	0.33
RAE-NWM	**2.91**	**0.70**	1.36	0.37	**1.14**	**0.28**

each candidate, the world model performs an eight-step rollout in the uncompressed visual token space conditioned on the historical observations. Instead of decoding to pixel space and using low-level metrics, we compute the DINO feature distance directly in token space between the predicted rollout and the goal image. We select the highest-scoring trajectory for execution. To quantify planning accuracy, we compute the Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) against the ground-truth trajectory. Table 2 also includes a deterministic DINO-token regression variant, which is further discussed in the ablation study.

Table 2 shows that RAE-NWM improves upon the primary baseline NWM on the SACSoN dataset, reducing the ATE to 2.91 and RPE to 0.70. This trend continues on the SCAND dataset. On the unstructured RECON dataset, the proposed model yields a competitive ATE of 1.36 and RPE of 0.37, outperforming end-to-end policies like GNM and NoMaD. High-frequency visual elements, such as the dense off-road vegetation in the RECON dataset, introduce stochastic textures that complicate environment modeling. In this short-horizon setting, NWM obtains lower ATE and RPE on RECON. This performance gap likely occurs because the VAE-based latent space preserves more texture-sensitive cues, while DINOv2 features emphasize semantic and structural information.

Table 3: Closed-loop image-goal navigation results in Habitat.

Method	SR (%) ↑	SPL (%) ↑
NoMaD [38]	16.67	15.13
OmniVLA [18]	36.67	34.78
NWM [5]	43.33	38.66
One-Step WM [36]	72.67	**69.10**
RAE-NWM	**78.95**	63.58

Table 4: Dynamics condition ablation studies on the SACSoN dataset.

Method	ATE ↓	RPE ↓
Simple Addition	3.34	0.79
MLP Fusion	3.54	0.83
Scheduled Gate	3.31	0.78
Learned Gate (ours)	**2.91**	**0.70**

5.4 Simulation in Habitat

To evaluate closed-loop control capabilities, we deploy RAE-NWM as a dynamics engine in the Habitat simulator using scenes from Matterport3D. We assess the model on the Image-Goal navigation task, where the agent navigates to a destination specified by a goal image. An episode (average length 8 meters) is considered successful if the agent stops within 1 meter of the target. At each environment step, the agent performs CEM-based trajectory optimization guided by the same token-space DINO feature distance used in Section 5.3, and executes the selected action accordingly. Table 3 shows that RAE-NWM achieves a Success Rate of 78.95%, outperforming existing visual navigation methods. Its SPL is slightly lower than that of One-Step WM, likely because one-step generation can evaluate longer planning trajectories more efficiently under a fixed computation budget, while RAE-NWM attains a higher SR with iterative dense-space rollouts.

5.5 Ablation Study

To study the impact of each component, we evaluate both long-horizon rollout quality and downstream planning accuracy. Sequential generation over a 16-second horizon serves as a stress test for temporal stability, while the planning task further measures action controllability.

Transition Modeling Objective. To isolate the effect of the generative transition formulation from the choice of dense DINO representation, we implement a deterministic DINO-token Regression (DINO-Reg) variant under the same first-person navigation protocol. This variant keeps the representation space and evaluation protocol unchanged, but replaces flow matching with direct MSE regression over future DINO tokens. As shown in Table 2, RAE-NWM obtains lower ATE on all datasets and lower RPE on SACSoN and SCAND, while DINO-Reg is marginally better on RECON RPE (0.36 vs. 0.37). This suggests that the transition objective contributes beyond the representation choice alone. The qualitative rollout in Figure 7 further shows that deterministic regression produces blurry predictions with ghosting artifacts, rather than stable frame-wise rollouts.

Fig. 7: Qualitative rollout visualization of DINO-Reg.

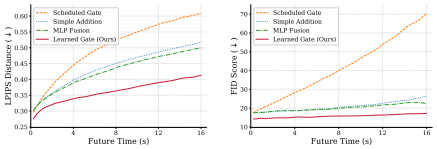


Fig. 8: Ablation on dynamics condition injection strategies.

Dynamics Conditioning Module. To investigate the optimal condition injection strategy during the continuous generative process, we compare our adaptive learned gate against simple addition, MLP fusion, and scheduled gate. As Figure 8 illustrates, standard injection strategies and the scheduled gate experience continuous degradation in both LPIPS and FID over extended horizons, indicating that rigid conditioning is less effective at preserving spatial structure. In contrast, the learned gate maintains the lowest LPIPS trajectory and a more stable FID curve. We further evaluate these variants in downstream planning on SACSoN (Table 4), where the learned gate achieves the lowest ATE of 2.91 and RPE of 0.70. These results show that adaptively modulating the injected dynamics representation reduces visual error accumulation and improves precise action controllability. Detailed conditioning designs and gate dynamics are provided in Section D of the Supplementary Material.

Representation Space and Architecture. To test the impact of the representation space, we substitute the DINOv2 encoder with the same frozen SD-VAE [29] used in NWM. As illustrated in Figure 9, while SD-VAE maintains a lower initial LPIPS due to its pixel-level reconstruction objective, it suffers from stronger degradation over extended horizons. The DINOv2-based model quickly surpasses it, exhibiting better long-term spatial stability and a superior final FID. We also remove the DDT head to evaluate the architectural design. Without the DDT head, the long-horizon LPIPS exceeds the SD-VAE baseline, suggesting that directly modeling high-dimensional DINOv2 representations remains difficult. Consistent with RAE [53], the DDT head effectively improves optimization in this dense token space, enabling the model to better exploit the geometric priors of DINOv2 during continuous rollouts.

6 Discussion and Limitations

While RAE-NWM excels in structured environments, semantic representations like DINOv2 tend to overlook high-frequency stochastic textures (e.g., grass in the RECON dataset). Figure 10 illustrates that this initial fidelity loss acts as a trade-off for long-horizon spatial stability. The NWM starts with a slightly lower LPIPS (0.230), but its limited structural grounding leads to rapid degradation,

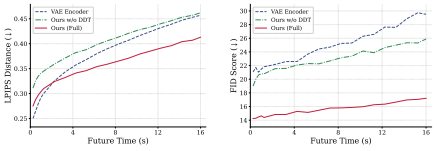


Fig. 9: Ablations on visual encoders and the DDT head.

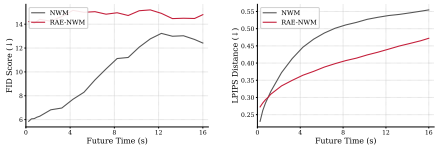


Fig. 10: Quantitative performance on the RECON dataset.

reaching 0.554 during sequential rollouts. RAE-NWM maintains a more stable trajectory, ending at 0.472. This indicates that semantic priors help maintain coherent geometric structure during dynamics generation, even though some texture-sensitive details are less faithfully preserved. We further measure static encode-decode fidelity on RECON without dynamics prediction, where SD-VAE and DINOv2+RAE obtain FID scores of 4.14 and 5.47, respectively, confirming the texture-fidelity cost of DINOv2+RAE.

Despite using a smaller generative backbone than NWM (350M vs. 1B parameters; 256 vs. 784 dynamics tokens), RAE-NWM achieves improved performance with lower dynamics-forward cost in our profiling (8.97 vs. 36.10 TFLOPs; 19.4 vs. 28.8 ms; 0.73 vs. 1.93 GB on one A800). Fully scale-matched training remains a limitation for isolating the effects of representation choice, architecture, and model capacity.

7 Conclusion

We present the Representation Autoencoder-based Navigation World Model (RAE-NWM), a generative navigation world model that learns action-conditioned transitions in the dense DINOv2 representation space. Built upon a frozen DINOv2 encoder and a frozen RAE decoder, RAE-NWM trains a CDiT-DH backbone to model continuous transition dynamics and introduces a time-driven gating mechanism to adaptively modulate kinematic conditioning along the probability flow. This design improves long-horizon rollout stability, preserves spatial geometric structure, and provides more reliable predictions for downstream planning. Extensive experiments on real-world navigation datasets and Habitat Image-Goal navigation demonstrate the effectiveness of dense representation-space world modeling. Future work will explore improving texture fidelity and accelerating iterative rollouts for more efficient planning.

Acknowledgments

This work was supported in part by Tsinghua-Toyota Joint Research Fund, in part by National Natural Science Foundation of China (Grant No. 62403269), and in part by Beijing Natural Science Foundation under grant number L233029.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
2. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
3. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
4. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.Y.: eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
5. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation World Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15791–15801. Nashville, TN, USA (June 11–15 2025)
6. Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: Proceedings of the International Conference on 3D Vision (3DV). pp. 667–676. Qingdao, China (October 10–12 2017)
7. Chen, H., Han, Y., Chen, F., Li, X., Wang, Y., Wang, J., Wang, Z., Liu, Z., Zou, D., Raj, B.: Masked autoencoders are effective tokenizers for diffusion models. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 267, pp. 8145–8171. PMLR, Vancouver, BC, Canada (July 13–19 2025)
8. Chen, J., Zou, D., He, W., Chen, J., Xie, E., Han, S., Cai, H.: DC-AE 1.5: Accelerating diffusion model convergence with structured latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19628–19637. Honolulu, HI, USA (October 19–23 2025)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). Virtual Event, Austria (May 3–7 2021)
10. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A survey of embodied AI: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence **6**(2), 230–244 (2022)
11. 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. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 50742–50768. New Orleans, LA, USA (December 10–16 2023)
12. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York, 2 edn. (2009)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009. New Orleans, LA, USA (June 19–24 2022)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. Las Vegas, NV, USA (June 27–30 2016)
16. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: International Conference on Learning Representations (ICLR). Kigali, Rwanda (May 1–5 2023)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30, pp. 6629–6640. Long Beach, CA, USA (December 4–9 2017)
18. Hirose, N., Glossop, C., Shah, D., Levine, S.: OmniVLA: An omni-modal vision-language-action model for robot navigation. arXiv preprint arXiv:2509.19480 (2025)
19. Hirose, N., Shah, D., Sridhar, A., Levine, S.: SACSoN: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters **9**(1), 49–56 (2024)
20. 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)
21. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Vuong, Q., Sanketi, P., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: OpenVLA: An open-source vision-language-action model. In: Proceedings of the 8th Conference on Robot Learning (CoRL). Proceedings of Machine Learning Research, vol. 270, pp. 2679–2713. PMLR, Munich, Germany (November 6–9 2025)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR). Banff, AB, Canada (April 14–16 2014)
23. Li, X., He, X., Zhang, L., Wu, M., Li, X., Liu, Y.: A comprehensive survey on world models for embodied AI. arXiv preprint arXiv:2510.16732 (2025)
24. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth Anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
25. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (ICLR). Kigali, Rwanda (May 1–5 2023)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205. Paris, France (October 2–6 2023)
28. Qin, H., Shen, W., Meng, Z., Li, X.: ESEM: A visual topological navigation method integrating edge semantic enhancement in challenging environments. In: 2025 IEEE 19th International Conference on Control & Automation (ICCA). pp. 25–31. Tallinn, Estonia (June 30–July 3 2025)

29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695. New Orleans, LA, USA (June 19–24 2022)
30. Shah, D., Eysenbach, B., Rhinehart, N., Levine, S.: Rapid exploration for open-world navigation with latent goal models. In: Proceedings of the 5th Conference on Robot Learning (CoRL). Proceedings of Machine Learning Research, vol. 164, pp. 674–684. PMLR, London, UK (November 8–11 2022)
31. Shah, D., Levine, S.: ViKiNG: Vision-based kilometer-scale navigation with geographic hints. In: Proceedings of Robotics: Science and Systems (RSS). New York City, NY, USA (June 2022). https://doi.org/10.15607/RSS.2022.XVIII.019
32. 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. London, UK (May 29–June 2 2023)
33. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: ViNT: A foundation model for visual navigation. In: Proceedings of the 7th Conference on Robot Learning (CoRL). Proceedings of Machine Learning Research, vol. 229, pp. 711–733. PMLR, Atlanta, GA, USA (November 6–9 2023)
34. Shang, Y., Jin, L., Ma, Y., Zhang, X., Gao, C., Wu, W., Li, Y.: LongScape: Advancing long-horizon embodied world models with context-aware MoE. arXiv preprint arXiv:2509.21790 (2025)
35. Shen, W., Gu, P., Qin, H., Meng, Z.: EffoNAV: An effective foundation-model-based visual navigation approach in challenging environments. IEEE Robotics and Automation Letters **10**(7), 6824–6831 (2025). https://doi.org/10.1109/LRA.2025.3572815
36. Shen, W., Meng, Z., Ma, J., Zhou, M., Xiang, D.: An efficient and multi-modal navigation system with one-step world model. arXiv preprint arXiv:2601.12277 (2026)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (ICLR). Virtual Event, Austria (May 3–7 2021)
38. 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. Yokohama, Japan (May 13–17 2024)
39. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of RGB-D SLAM systems. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 573–580. Vilamoura, Algarve, Portugal (October 7–12 2012)
40. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D.S., Maksymets, O., et al.: Habitat 2.0: Training home assistants to rearrange their habitat. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 251–266. Virtual (December 6–14 2021)
41. Tang, H., Xie, C., Bao, X., Weng, T., Li, P., Zheng, Y., Wang, L.: UniLiP: Adapting CLIP for unified multimodal understanding, generation and editing. arXiv preprint arXiv:2507.23278 (2025)
42. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O.J., Harmsen, J., Steiner, A., Zhai, X.: SigLIP 2: Multilingual vision-language encoders with

improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
43. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5294–5306. Nashville, TN, USA (June 11–15 2025)
44. Wang, S., Tian, Z., Huang, W., Wang, L.: DDT: Decoupled diffusion transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 40633–40642. Denver, CO, USA (June 3–7 2026)
45. Xia, Z., Lu, Y., Li, X., Xu, Y., Chen, Y.: Cloning deterministic worlds: The critical role of latent geometry in long-horizon world models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 32602–32612. Denver, CO, USA (June 3–7 2026)
46. Yang, J., Li, T., Fan, L., Tian, Y., Wang, Y.: Latent denoising makes good tokenizers. In: International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil (April 23–27 2026)
47. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15703–15712. Nashville, TN, USA (June 11–15 2025)
48. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 128940–128966. Vancouver, BC, Canada (December 10–15 2024)
49. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–18 (2025). https://doi.org/10.1109/TPAMI.2025.3613256, early access
50. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595. Salt Lake City, UT, USA (June 18–22 2018)
51. Zhang, W., Tang, P., Zeng, X., Man, F., Yu, S., Dai, Z., Zhao, B., Chen, H., Shang, Y., Wu, W., et al.: Aerial world model for long-horizon visual generation and navigation in 3D space. arXiv preprint arXiv:2512.21887 (2025)
52. Zhang, Z., Zhang, H., Huang, K., Shi, C., Lu, H.: Efficient image-goal navigation with representative latent world model. arXiv preprint arXiv:2511.11011 (2025)
53. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. In: International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil (April 23–27 2026)
54. Zhou, G., Pan, H., LeCun, Y., Pinto, L.: DINO-WM: World models on pre-trained visual features enable zero-shot planning. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 267, pp. 79115–79135. PMLR, Vancouver, BC, Canada (July 13–19 2025)
55. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 3357–3364. Singapore (May 29–June 3 2017)