Pondering the Way: Spatial-perceiving World Action Model for Embodied Navigation

Hong Chen[1](✉), Daqi Liu[2⋆](✉), Zehan Zhang[2⋆](✉), Haiguang Wang[2], Tianhao Lu[1], Longfei Yan[3], Haiyang Sun[2], Fangzhen Li[2], Hongwei Xie[2], Bing Wang[2], Guang Chen[2], Hangjun Ye[2⋆⋆](✉), and Yihua Tan[1⋆⋆](✉)

[1] Huazhong University of Science and Technology, Wuhan, China
{hongc, yhtan}@hust.edu.cn
[2] Xiaomi EV, Beijing, China
liudaqikk@gmail.com, zehanzhang@126.com, yehangjun@gmail.com
[3] Zhejiang University, Hangzhou, China

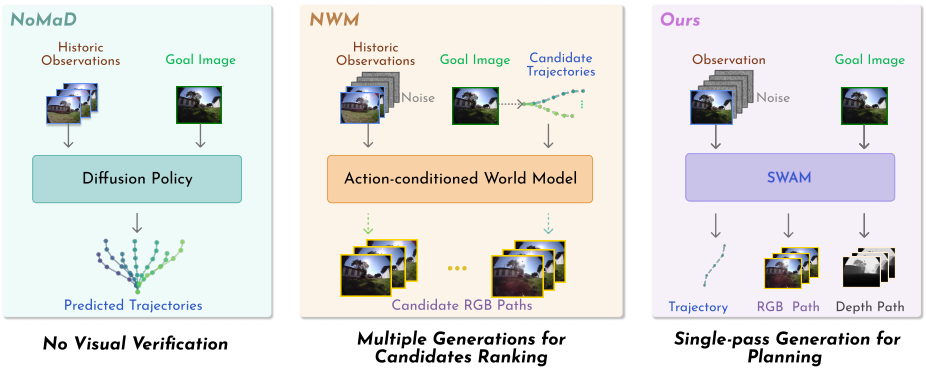

Fig. 1: Comparison of different paradigms for visual navigation. **Left:** Nomad predicts trajectories from historical RGB frames and a goal image. **Middle:** NWM generates future RGB videos conditioned on trajectories. **Right (Ours):** Our model jointly predicts trajectories and RGBD paths from the current RGB observation and the goal image in one forward pass, enabling efficient planning.

Abstract. Existing world model-based planners for visual navigation typically follow a verification-centric paradigm, decoupling goal intent from trajectory synthesis. This approach suffers from candidate dependence, heavy computational overhead, and inconsistencies between sampled actions and predicted visuals. To address these issues, we propose SWAM (**S**patial-perceiving **W**orld **A**ction **M**odel), a task-centric joint observation-action generation framework. Given start and goal RGB observations, SWAM performs single-pass inference to simultaneously generate intermediate RGB-D sequences and corresponding action trajectories, promoting goal-consistent trajectory generation and improved spatial feasibility. SWAM leverages depth pseudo-labels during training to

⋆ Project leaders.
⋆⋆ Corresponding authors.

internalize spatial priors, but it requires only monocular RGB input at inference time by using depth estimation model. We further introduce a visual-guided action refinement module and a trajectory-scale regularization loss to enforce fine-grained alignment between motion and visual cues while stabilizing predictions across varying distances. Extensive experiments show that SWAM significantly outperforms state-of-the-art two-stage planners in success rate, trajectory accuracy, and inference efficiency, while demonstrating robust zero-shot generalization to unseen environments.

Keywords: Visual Navigation · World Action Model · Joint Generation

1 Introduction

Visual goal-conditioned navigation [4, 51] is a core challenge for intelligent systems operating in the physical world. The task demands that an agent, given a visual goal and its current observation, must infer where to move next, which paths are traversable, and when the goal has been successfully reached. This requires tight integration of visual perception, spatial reasoning, and decision-making [8, 13].

Direct online policy methods learn observation-to-action mappings [29,31,32, 36, 37, 41]. While inference-efficient, they lack explicit visual horizons and struggle with error recovery. Offline world model-based planners [2, 45] adopt a two-stage pipeline: sample candidate action sequences from an external policy [6,36], rollout visual trajectories via an action-conditioned video predictor, and select the trajectory best matching the goal. While this enables simulation-evidence-based evaluation, its modular design constrains the world model to verification rather than trajectory synthesis, leading to three limitations: (i) candidate dependence, decision quality is bounded by candidate coverage; (ii) computational inefficiency, exhaustive sampling and rollout incur prohibitive inference costs, especially in complex or longer planning settings; (iii) geometric inconsistency, absence of explicit spatial constraints can produce abrupt viewpoint changes or infeasible trajectories, reducing the reliability of simulation evidence.

We argue that these limitations stem from decoupling the path proposal and environment prediction. Verification-centric pipelines treat the world model [28, 40, 43] as a passive evaluator, decoupling goal intent from trajectory generation and relegating spatial reasoning to implicit visual patterns, which neglects the core requirement of navigation.

To address this, we propose a Spatial-perceiving World Action Model (SWAM), a joint observation–action generation framework. Given a start RGB observation and a goal RGB observation, our model performs single-pass inference to jointly generate an intermediate RGBD observation sequence and the corresponding action trajectory. Unlike unidirectional world models [2, 54], our approach co-generates "how to act" and "what to perceive" within a unified process. This approach jointly constrains actions and observations throughout

generation, ensuring goal alignment, temporal consistency, and spatial feasibility. Unlike verification-centric world models, SWAM directly synthesizes control trajectories through joint latent denoising. SWAM offers three key advantages: (1) goal alignment is maintained throughout generation, producing structured, target-directed paths; (2) inference cost is reduced by eliminating candidate expansion and repeated evaluation; (3) explicit injection of spatial cues enhances the constraints on actions.

Technically, we leverage DepthAnything v3 [21] to provide per-frame depth pseudo-labels during training, leveraging spatial priors while requiring only monocular RGB at test time. Estimated depth supplies sufficient navigable cues, avoiding reliance on scarce ground-truth depth. We further introduce a Visual-Guided Action Refinement (VGAR) module and Trajectory-Scale Regularization (TSR) loss to enhance the alignment between predicted actions and generated visual cues, and to stabilize trajectory predictions over various distances.

Experiments show that SWAM outperforms strong policy-based and two-stage baselines in trajectory error, goal success rate, visual fidelity, and practical inference efficiency, while demonstrating more stable long-distance and cross-dataset behavior. In conclusion, our contributions are threefold:

- We introduce SWAM, a joint observation–action generation framework for visual navigation, replacing candidate-based pipelines;
- We propose a visual-guided action refinement and trajectory-scale regularization loss to better leverage visual cues for better action prediction and reduce trajectory drift error;
- We validate SWAM across multiple datasets, showing improved accuracy, success rate, visual quality, inference efficiency, and robust cross-scene generalization.

2 Related Work

2.1 Visual Goal-Conditioned Navigation

Visual goal-conditioned navigation aims to learn policies that guide an agent to reach a target location specified by a goal observation. Early approaches formulate this problem as learning a mapping from the current observation and a goal image to control actions, typically using reinforcement learning or imitation learning [11, 27, 56]. These methods learn deterministic policies that directly predict actions from visual observations and have demonstrated strong performance in simulated indoor environments. Subsequent work improves policy learning through representation learning and memory mechanisms, enabling agents to reason over partial observations and long-horizon dependencies [24,25]. However, deterministic policy learning often struggles with the multi-modality of navigation behaviors, where multiple valid trajectories can reach the same goal. To address this limitation, recent works explore generative formulations of navigation policies. In particular, diffusion-based policies have been proposed to model distributions over action trajectories, enabling more diverse and robust

navigation strategies [6, 36]. While these approaches improve the expressiveness of goal-conditioned policies, they typically generate action sequences alone, without explicitly modeling the intermediate visual states along the trajectory. In contrast, our approach formulates navigation as joint video–action generation conditioned on the goal observation, allowing the model to explicitly capture the visual transitions connecting the start and goal states.

2.2 Action-Conditioned World Models for Planning and Navigation

World models [1, 14, 20] aim to learn predictive models of environment dynamics that allow agents to plan by imagining future observations under candidate actions. Early works, such as visual model predictive control, predict future images conditioned on candidate action sequences and optimize actions by minimizing perceptual distance to a goal observation [7, 16]. Recent advances in generative modeling [17, 22, 34] have significantly improved the capacity of world models. A number of works adopt video prediction or video diffusion models to model environment dynamics for planning [12]. More recently, large-scale generative models have been used to construct action-conditioned video diffusion world models [23,39,52] capable of generating realistic imagined rollouts [2,9,45,50,54]. These predicted trajectories can then be used for planning via trajectory ranking or search. Despite their strong generative ability, these pipelines generally follow an action-first, verify-later paradigm, where candidate actions are first sampled from policies or search algorithms and then evaluated using the world model. As a result, planning performance depends heavily on the coverage of candidate trajectories and the available rollout budget, which can become computationally expensive for long-horizon navigation tasks. In contrast, our method directly generates coherent action–observation trajectories conditioned on the goal in a single generative process, avoiding candidate trajectory sampling and improving planning efficiency.

2.3 Joint Video Action Modeling for Embodied Intelligence

Recent advances in embodied AI emphasize the joint modeling of observations and actions to capture the causal dynamics of agent-environment interactions. By unifying these modalities within generative world models, researchers have significantly improved agent planning, controllability, and generalization.

Navigation-oriented world models have transitioned from basic future observation predictors to spatially grounded reasoning systems. Models such as DINO-WM [53] and WoMaP [48] leverage pretrained visual features and structured modeling to enhance open-vocabulary localization and zero-shot planning. Further scaling perception-action integration, World-in-World [49] and UniDrive-WM [42] explore closed-loop learning and unified perception-planning in driving scenarios. However, these methods primarily treat the world model as an environment simulator, often decoupling environment dynamics from the explicit generation of executable action sequences.

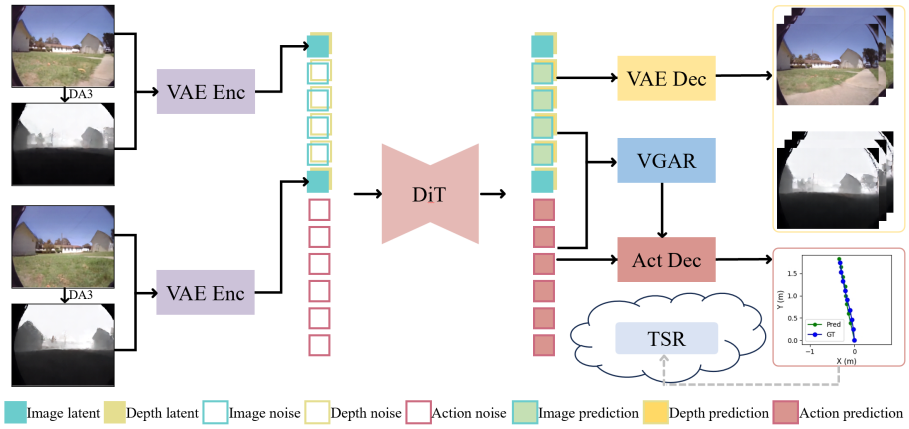

Fig. 2: Overview of SWAM. We extend the pretrained Diffusion Transformer (DiT) of CogVideoX [46] by conditioning it on start and goal frame latents, and fine-tune it to jointly produce intermediate RGB-D frame latents and associated action tokens. DepthAnything V3 [21] is used to predict pseudo-depth maps for the start and goal frames. The Visual-Guided Action Refinement (VGAR) module further refines the predicted actions via cross-attention with RGBD latents. During training, losses are imposed both on the diffusion denoising process and our proposed Trajectory-Scale Regularization (TSR) loss to ensure local plausibility and global trajectory consistency.

A parallel trend focuses on joint video-action generation, where models like VideoVLA [33], CoVAR [44], and others [3, 5, 19] demonstrate that unified architectures can excel in robotic manipulation. Recent works such as World Action Models [47] shows that action-aware world models can serve as zero-shot policies. While PAD [10] shares our goal of joint modeling, these existing approaches predominantly target short-horizon manipulation or local workspace interactions. They frequently lack the long-horizon spatiotemporal priors and explicit geometric grounding essential for complex navigation tasks. More recently, Aether [55] combines geometric reconstruction and generative modeling to improve spatial understanding and prediction. Although navigation-related scenarios are included, navigation-specific planning metrics and goal-conditioned trajectory generation are not explicitly evaluated.

In contrast to the aforementioned research, SWAM addresses various-distance planning through a unified perception-action generative process with spatial perceiving.

3 Method

In this section, we present the proposed SWAM framework for goal-conditioned visual navigation, including the unified observation-action generation architecture, the Visual-Guided Action Refinement (VGAR) module, the Trajectory-Scale Regularization (TSR) loss, and the training strategy built upon a pre-

trained video diffusion backbone. More detailed information is shown in the Appendix.

3.1 Problem Formulation

We consider goal-conditioned visual navigation, where an agent is given a start observation and a goal observation, and is required to generate a feasible trajectory connecting them. Let $o_0 = (I_0, D_0)$ and $o_G = (I_G, D_G)$ denote the start and goal observations, where I represents an RGB image and D denotes the corresponding depth representation.

Our objective is to jointly generate a horizon-N sequence of future observations ($\{o_n\}_{n=1}^N$) and actions ($\{a_n\}_{n=1}^N$, where $a_n = (\Delta x_n, \Delta y_n)$) denotes the planar motion of the robot in its local coordinate frame. Specifically, we learn the conditional distribution

$$p_\theta \left(\{o_n\}_{n=1}^N, \{a_n\}_{n=1}^N \mid o_0, o_G \right).$$

During training, depth representations are obtained using an off-the-shelf monocular depth estimator and are used as auxiliary geometric supervision to provide spatial grounding for the generative process.

3.2 Overall Architecture

Fig. 2 illustrates the overall framework. We adopt a latent diffusion model that jointly generates (i) an RGBD observation sequence and (ii) the corresponding action sequence. This joint modeling encourages temporal-spatial consistency: actions are optimized to be compatible with the predicted visual evolution, and the generated visual evolution is conditioned on the action plan.

Latent Representation and Tokenization. Given each observation denoted as $o_n = (I_n, \hat{D}_n)$, we employ the frozen 3D variational autoencoder (VAE) of CogvideoX with spatial-temporal convolutions to encode the RGB image and depth map into latent feature maps, separately: RGB latent $z_n^i \in \mathbb{R}^{h \times w \times d}$ and depth latent $z_n^d \in \mathbb{R}^{h \times w \times d}$, where $h \times w$ is the latent spatial resolution and d is the feature dimension. We represent an action sequence $\{a_n\}_{n=1}^N$ as a token matrix $x^a \in \mathbb{R}^{N \times D}$ by projecting each 2D action into the same embedding dimension using a learnable MLP. This enables unified modeling over visual and action tokens.

Multimodal Sequence Construction. We construct one unified token sequence that contains conditions and denoised targets. Specifically, we define the clean target variables (to be diffused) as: intermediate RGB tokens $\{z_n^i\}_{n=1}^N$, intermediate depth tokens $\{z_n^d\}_{n=1}^N$, and action tokens x^a. We concatenate them into a single sequence X_t and treat the start/goal latents as conditioning information rather than part of X_t: $c = (z_0^i, z_0^d, z_G^i, z_G^d)$, where t represents the denoising step.

Denoising Process. We adopt a DDPM-style forward noising process on the unified sequence. In the denoising step t, the diffusion transformers (DiTs) [26]

$\mathcal{F}_\theta$ iteratively denoise the sequence over T steps. At denoising step t, the model predicts the noise residual ϵ_θ given the current noisy sequence X_t and goal conditioning:

$$X_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(X_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(X_t, z_0^i, z_0^d, z_G^i, z_G^d, t) \right) + \sigma_t z, \tag{1}$$

where α_t and $\bar{\alpha}_t$ are diffusion schedule parameters, σ_t is the sampling variance, and $z \sim \mathcal{N}(0, I)$ (omitted when $t = 0$). After denoising completes, RGB and depth latents are decoded to image space using the VAE decoder, while action tokens are projected to planar actions through a lightweight network f with the proposed VGAR module as described in Sec 3.3.

3.3 Visual-Guided Action Refinement

Since the final action readout is less sensitive to local geometric cues, we introduce a lightweight Visual-Guided Action Refinement (VGAR) module that injects generated RGBD evidence into action tokens right before decoding, without modifying the joint diffusion backbone. Let X_v and X_a denote the final-layer visual and action tokens from the diffusion Transformer. VGAR refines X_a via a gated residual cross-attention:

$$C = \mathrm{CrossAttn}\big(\mathrm{LN}(X_a),\ \mathrm{LN}(X_v),\ \mathrm{LN}(X_v)\big), \tag{2}$$

$$\Delta X_a = G \odot W(C), \quad G = \sigma\big(\mathrm{MLP}([\mathrm{LN}(X_a);\ C])\big), \tag{3}$$

$$X_a' = X_a + \Delta X_a, \tag{4}$$

where LN is LayerNorm, $[\cdot;\cdot]$ denotes feature concatenation, W is a learnable projection, and σ is the sigmoid function. The gate G controls how much visual evidence flows into the action representation. An action head then decodes the refined tokens X_a' into action sequences.

3.4 Joint-Training Objective

Diffusion Loss. We train with the standard DDPM denoising objective [17,35]:

$$\mathcal{L}_{\mathrm{DDPM}} = \mathbb{E}_{t, X_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\, X_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\ \mathbf{c},\ t\big) \right\|_2^2 \right], \tag{5}$$

where $\mathbf{c} = (z_0^i, z_0^d, z_G^i, z_G^d)$ denotes the conditioning set (image, depth, and goal tokens), X_0 is the clean latent sequence (RGB, depth, and action tokens), and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$ is the cumulative noise schedule.

Trajectory-Scale Regularization. We observe that diffusion loss alone on actions leads to compounding error when actions are integrated into a trajectory, with drift growing more severe over long horizons. To counteract this, we

introduce a Trajectory-Scale Regularization (TSR) loss that directly supervises the cumulative displacement:

$$\mathcal{L}_{\text{TSR}} = \frac{\|\hat{p}_N - p_G\|_2}{N}, \quad \hat{p}_N = \sum_{n=1}^{N} \hat{a}_n, \tag{6}$$

where $\hat{p}_N$ is the predicted endpoint obtained by integrating denoised actions and p_G is the ground-truth goal position in the local coordinate frame. The $1/N$ normalization ensures stability across trajectories of varying length.

Training Objective. The final objective combines diffusion denoising with the trajectory regularizer:

$$\mathcal{L} = \mathcal{L}_{\text{DDPM}} + \lambda_{\text{TSR}} \mathcal{L}_{\text{TSR}}. \tag{7}$$

Among sequences with similar local denoising error, TSR favors those that preserve the correct global displacement, turning long-distance drift into a directly learnable signal. This improves stability and generalization under varying trajectory lengths and limited-data regimes.

3.5 Model Initialization and Architectural Adaptation

To preserve the rich spatiotemporal priors of the pretrained video generation model while avoiding distribution shifts caused by newly introduced parameters, we initialize all newly added modules with zero weights. This strategy ensures that, at the beginning of training, the model behaves identically to the original pretrained generator and produces high-quality RGB predictions based on its learned priors. During training, gradients gradually update the new parameters, allowing the model to progressively learn the coupling between RGB observations, depth signals, and actions in a data-driven manner. Such progressive adaptation stabilizes training and enables seamless integration of additional modalities without disrupting the pretrained feature space.

4 Experiments

In this section, we present a comprehensive evaluation of our Spatial-perceiving World Action Model (SWAM). We first describe the experimental setup in Section 4.1, including datasets, baseline methods, and evaluation metrics. Section 4.2 presents the main quantitative results, followed by zero-shot generalization performance in Section 4.4, demonstrating the superiority of our proposed framework. We then provide qualitative analysis in Section 4.3 to illustrate the visual and trajectory consistency of our approach (additional visualizations are provided in the Appendix). Finally, Section 4.5 conducts ablation studies to validate the effectiveness of key components in our framework.

4.1 Experimental Setup

Datasets. We conduct a comprehensive evaluation of our method on four benchmark navigation datasets. RECON [30] comprises large-scale outdoor trajectories with extended horizons, thereby testing the model's capacity for long-term planning in complex real-world environments. SCAND [18] encompasses both indoor and outdoor scenes with dynamic obstacles and social agents, presenting challenges in reasoning about real-time interactions and heterogeneous scene structures. TartanDrive [38] evaluates performance in off-road driving scenarios characterized by pronounced forward-motion biases and irregular terrain, which necessitates adaptation to constrained and non-uniform action distributions. Finally, HuRoN [15] provides an indoor setting with dynamic human interactions, and we utilize it for evaluating the zero-shot generalization to previously unseen social dynamics. For all datasets, action signals are derived from consecutive robot poses and normalized to a unified scale across different embodiments, ensuring comparability of policy outputs across diverse navigation contexts. We follow the same data preprocessing, trajectory segmentation, and evaluation protocols established by NWM [2] during evaluation to ensure methodological consistency and fair comparison.

Baselines. We evaluate our approach against representative direct policy methods, including GNM [31] and the diffusion-based navigation policy NoMaD [36]. Additionally, we consider world-model-based planners by integrating NWM [2] with NoMaD [36] to generate candidate trajectories, which are subsequently ranked based on perceptual similarity. We evaluate NWM using varying numbers of sampled candidate trajectories ($N \in \{2, 4, 8, 16\}$) for ranking to demonstrate how candidate diversity influences final planning performance. This ensures the $\times N$ notation reflects the inference-time sampling budget rather than model scaling.

Since SWAM is initialized from the publicly available CogVideoX [46] backbone, we additionally construct a strong CogVideoX-based baseline to isolate the contribution of the proposed navigation-specific designs for fair comparisons. Specifically, we extend CogVideoX to jointly generate future observations and action sequences under the same training protocol, conditioning strategy, and pretrained initialization used by SWAM. The baseline predicts future RGB observations and corresponding actions conditioned on the start and goal frames, but does not incorporate depth modeling, trajectory-scale regularization, or visual-guided action refinement. This baseline enables an ablation-style evaluation, isolating the contribution of our proposed components while maintaining a consistent backbone architecture or pretraining.

Evaluation Metrics. We evaluate navigation performance along three complementary dimensions: trajectory accuracy, goal-reaching success, and video generation quality. Unless otherwise specified or indicated for the meters metric, trajectory accuracy is quantified using Absolute Trajectory Error (ATE) and Relative Pose Error (RPE), with all metrics reported in unit grids for consistency with NWM [2]. Goal-reaching performance is measured via success@(τ) at multiple distance thresholds ($\tau \in 1.0, 0.5, 0.25$ grids), capturing both coarse and

Table 1: Trajectory error (ATE/RPE) across datasets. Lower values are better. **Bold** indicates best performance; underline indicates second best.

Method	RECON		SCAND		TartanDrive		Time(s)
	ATE	RPE	ATE	RPE	ATE	RPE	
GNM	1.85	0.54	2.18	0.61	6.68	1.63	0.12
NoMaD	1.90	0.53	2.36	0.52	6.66	1.36	0.21
NWM+NoMaD (×2)	1.87	0.52	2.41	0.51	6.45	1.32	31.37
NWM+NoMaD (×4)	1.82	0.52	2.42	0.49	6.40	1.32	63.83
NWM+NoMaD (×8)	1.74	0.50	2.37	0.48	6.31	1.31	118.32
NWM+NoMaD (×16)	1.53	0.49	2.18	0.46	6.23	1.30	245.98
CogVideoX (Joint)	2.09	0.73	2.25	0.67	4.90	1.07	14.12
Ours	**0.93**	**0.43**	**1.15**	**0.34**	**1.55**	**0.68**	16.91

fine-grained navigation success rate. To assess the quality of the generated video paths, we compute PSNR, SSIM, and LPIPS between predicted and ground-truth RGB frames. All metrics are evaluated on the same standardized subsets as NWM, and reported results correspond to the average over three independent runs. Inference time is measured per episode without parallelization, providing a realistic assessment of runtime efficiency.

4.2 Main Results

We report the main experimental results that validate the effectiveness of our SWAM approach. Further related results can be found in the appendix.

Trajectory Accuracy. Tab. 1 shows ATE/RPE across datasets. SWAM achieves substantially lower trajectory errors than all baselines with outperforming NWM + NoMaD (×16) by 39.2% (RECON), 47.2% (SCAND), and 75.1% (TartanDrive) in ATE. Notably, SWAM requires only single-pass inference (16.91s/sample), while NWM+NoMaD (×16) requires exhaustive rollout (245.98s/sample), demonstrating superior efficiency–accuracy tradeoff. Notably, CogVideoX achieves the second-best ATE on TartanDrive, significantly outperforming NWM-based methods. These results indicate that joint RGB-action generation provides useful cues for trajectory prediction. However, CogVideoX's performance remains far inferior to our full model (1.55), and its RPE is notably worse (1.07 vs. 0.68), indicating that visual coherence alone is insufficient for precise navigation. Our approach builds upon this insight by explicitly coupling action prediction with RGBD generation and spatial guidance, yielding both geometrically feasible trajectories and visually consistent rollouts.

Goal-reaching Performance. Fig. 3 reports success@(τ) curves across thresholds. SWAM shows particularly strong gains at strict thresholds ($\tau = 0.25$), achieving a 2.1× higher success than NWM+NoMaD (×16) on RECON. This

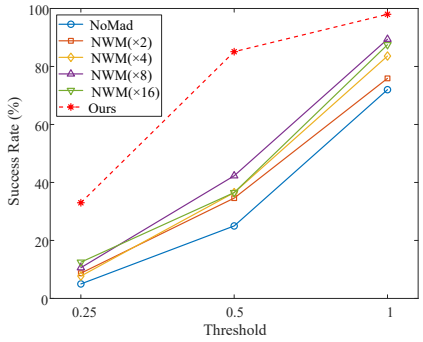

Fig. 3: Success@(τ) curves across thresholds on RECON. SWAM achieves substantially higher success rates, especially at strict thresholds ($\tau = 0.25$ and $\tau = 0.5$).

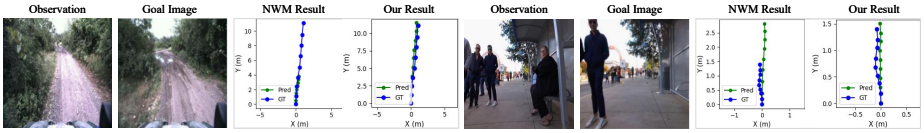

Fig. 4: Robustness to path length. SWAM maintains accurate trajectory scaling across distances, while NWM exhibits obvious trajectory scale errors.

indicates that SWAM enhances final-positional precision, which closely aligns with navigation planning.

Video Generation Quality. Tab. 2 summarizes video generation metrics. SWAM achieves state-of-the-art results across all datasets. We attribute the improved visual quality to two factors. First, spatial grounding via predicted depth provides strong geometric constraints that regularize the generation process, reducing visual artifacts and improving structural consistency across frames. This is particularly evident on TartanDrive, where our method achieves the highest PSNR (18.11) and SSIM (0.532), indicating that depth-aware generation produces more realistic terrain and obstacle appearances. Second, more accurate action prediction directly translates to better observation generation: since our actions are geometrically feasible and closely aligned with ground-truth trajectories, the corresponding RGBD frames naturally exhibit higher fidelity to the actual observations. This coupling ensures that improved action accuracy provides more realistic visual conditioning, which in turn refines future action predictions.

4.3 Qualitative Analysis

We provide qualitative analysis across different scenes and motion conditions. For NWM, we visualize the trajectory with the best trajectory ranking score to ensure a fair comparison. Fig. 4 evaluates robustness under nearly straight-line motion with varying trajectory lengths. SWAM maintains accurate trajec-

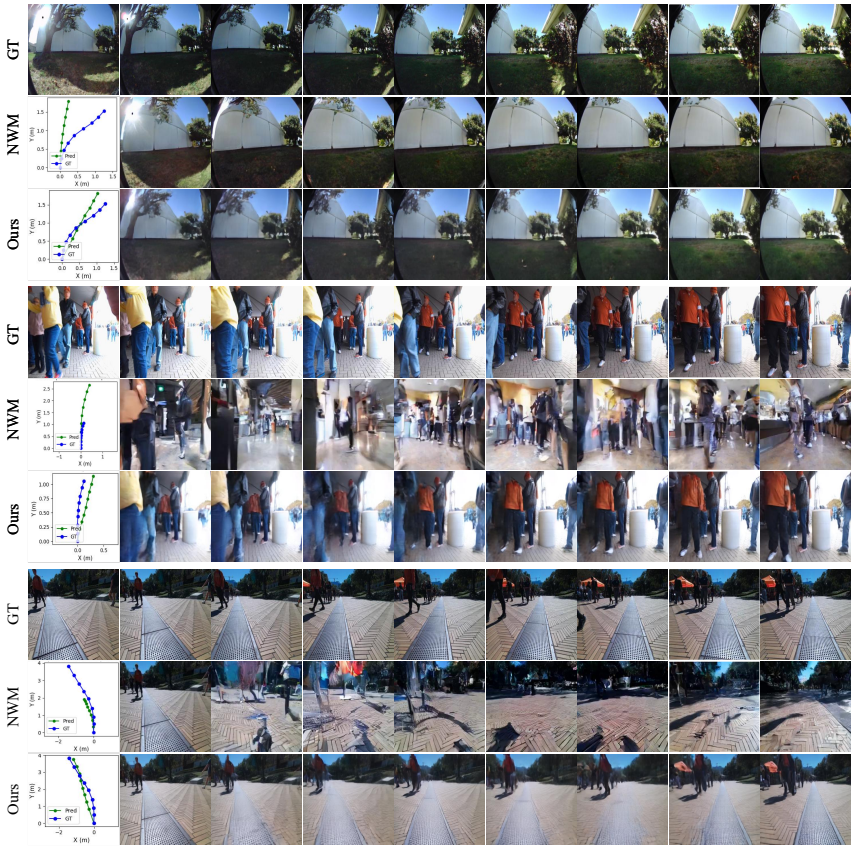

Fig. 5: Qualitative comparison between SWAM and NWM. SWAM generates trajectories closer to ground truth with consistent video generation, while NWM struggles in such challenging scenarios.

tory scaling across distances (e.g., $\sim$10 m and 1.5 m in two examples), whereas NWM exhibits scale drift that grows with the prediction horizon, indicating that SWAM effectively mitigates scale accumulation errors over varying distances. Fig. 5 compares trajectory and video generation across diverse cases. SWAM correctly captures turning trajectories where NWM shows a strong straight-motion bias. In complex open environments, SWAM preserves both trajectory accuracy and spatiotemporal consistency in generated observations, while NWM gradually loses contextual information, leading to mode collapse, hallucinated frames, and large trajectory errors. Overall, SWAM produces trajectories that closely follow the ground truth with minimal drift while maintaining temporally coherent video generation, whereas NWM struggles to jointly maintain geometric accuracy and visual consistency in challenging scenarios. These results demon-

Table 2: Video generation (PSNR/SSIM/LPIPS). Higher PSNR/SSIM and lower LPIPS are better. **Bold** indicates best performance; underline indicates second best.

Method	RECON			SCAND			TartanDrive		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NWM+NoMaD (×2)	11.84	0.360	0.470	11.66	0.360	0.476	13.75	0.276	0.402
NWM+NoMaD (×4)	14.14	0.435	0.364	11.85	0.367	0.468	13.84	0.277	0.427
NWM+NoMaD (×8)	14.47	0.439	0.351	11.79	0.363	0.470	13.64	0.278	0.390
NWM+NoMaD (×16)	14.45	0.438	0.350	11.95	0.368	0.471	13.22	0.284	0.378
CogVideoX (Joint)	16.65	0.633	0.366	15.04	0.601	0.351	17.20	0.524	**0.330**
Ours	**17.31**	**0.653**	**0.328**	**16.11**	**0.619**	**0.325**	**18.11**	**0.532**	0.335

Table 3: Zero-shot generalization to HuRoN (ATE/RPE). Lower values are better. **Bold** indicates best performance.

Method	ATE	RPE
NWM+NoMaD (×16) (trained on HuRoN)	3.73	0.95
Ours (zero-shot, no HuRoN training)	**2.94**	**0.85**

strate that SWAM better preserves the coupling between action dynamics and visual observations, leading to more reliable predictions.

4.4 Zero-Shot Generalization

We further evaluate the generalization ability of SWAM under unseen environments. For this evaluation, we conduct zero-shot transfer experiments on the HuRoN dataset without any training or fine-tuning. Tab. 3 shows the results. Despite never being trained on HuRoN, our model achieves an ATE of 2.94 and an RPE of 0.85, i.e., 21.2% lower ATE than NWM+NoMaD (×16), which was explicitly trained on HuRoN. This result demonstrates the strong cross-domain generalization capability of SWAM across environments with different layouts and visual characteristics.

4.5 Ablation Studies

To analyze the contribution of each component in our framework, we conduct ablation studies on three datasets: RECON, SCAND, and TartanDrive, using CogVideoX as the baseline. This model supports joint video-action generation, allowing us to fairly evaluate each enhancement while keeping the generative backbone consistent. Tab. 4 reports trajectory accuracy (ATE/RPE) and video generation quality (PSNR), providing a comprehensive view of how each module affects navigation precision and visual fidelity.

Effect of Additional Depth Modality. Adding depth modality provides explicit spatial perception cues for video generation and trajectory prediction. With depth incorporated, trajectory accuracy improves significantly over the baseline.

Table 4: Ablation study on RECON, SCAND, and TartanDrive. Lower ATE/RPE and higher PSNR are better. **Bold** indicates best performance; underline indicates second best.

Module			RECON			SCAND			TartanDrive		
Depth	TSR	VGAR	ATE↓	RPE↓	PSNR↑	ATE↓	RPE↓	PSNR↑	ATE↓	RPE↓	PSNR↑
Baseline			2.09	0.70	16.65	2.25	0.67	15.04	4.90	1.07	17.20
✓			1.70	0.47	**17.54**	2.27	0.51	16.05	4.61	0.97	17.93
	✓		2.06	0.69	16.37	1.63	0.45	15.44	2.63	0.94	17.17
✓	✓		1.01	0.45	17.30	**1.12**	**0.30**	**16.13**	1.94	**0.53**	18.12
✓	✓	✓	**0.94**	**0.43**	17.31	1.15	0.34	16.11	**1.55**	0.68	**18.15**

On RECON, ATE decreases from 2.09 to 1.70 and RPE from 0.70 to 0.47. Similar improvements occur on TartanDrive (ATE 4.90 → 4.61) and SCAND (RPE 0.67 → 0.51). These results indicate that depth enhances geometric understanding, yielding more physically consistent and reliable trajectories.

Effect of Trajectory-scale Regularization Loss. The Trajectory-scale Regularization (TSR) loss enforces endpoint alignment with navigation goals, leading to consistent improvements across datasets. As shown in Tab. 4, TSR alone improves over the baseline, reducing ATE from 2.09 to 2.06 on RECON and from 2.25 to 1.63 on SCAND, with a notable RPE reduction on SCAND (0.67 → 0.45). On TartanDrive, TSR substantially decreases ATE from 4.90 to 2.63, highlighting its effectiveness in long-horizon trajectory stabilization. When TSR is combined with Depth, performance is further improved compared to Depth only, achieving 1.70 → 1.01 ATE on RECON and 2.27 → 1.12 on SCAND, while also reducing TartanDrive ATE from 4.61 to 1.94, demonstrating consistent complementary gains in both indoor and driving scenarios.

Effect of Visual-Guided Action Refinement. The Visual-Guided Action Refinement (VGAR) module refines predicted trajectories using visual-spatial constraints from RGBD tokens. With VGAR, both ATE and RPE improve on most datasets; for TartanDrive, ATE reaches 1.55 and RPE 0.68, demonstrating effective trajectory refinement with a lightweight network.

Overall, the ablation studies confirm that each component meaningfully contributes to performance: depth prediction provides spatial grounding, TSR loss ensures goal-aligned trajectories, and VGAR further corrects long-range predictions. Their combination enables robust perception-action coupling, yielding accurate and physically plausible trajectory generation.

4.6 Failure Cases and Limitation Discussions

Despite its robustness, SWAM encounters challenges in scenarios that include ambiguous weeds and sharp turns, as illustrated in Fig. 6.

The primary failure mode occurs in scenarios involving ambiguous weeds, where the model frequently misidentifies traversable weeds as solid obstacles.

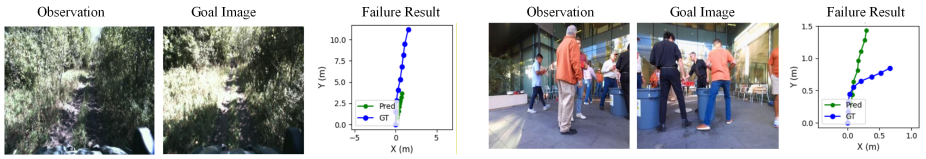

Fig. 6: Qualitative results of typical failure scenarios.

This perceptual ambiguity, stemming from the lack of semantic-aware traversability reasoning, leads to erroneous avoidance behaviors and accumulated trajectory drift. Another limitation arises in sharp-turn scenarios, where the current planar displacement based action representation fails to capture the rapid heading adjustments required for high-curvature paths. Because the framework does not explicitly encode orientation dynamics, the agent tends to deviate from intended sharp-turn trajectories.

Future work will investigate integrating semantic scene understanding into the model to enable semantic-aware traversability reasoning, as well as developing richer action representations that jointly model position and orientation for more robust long-horizon navigation. Moreover, extending prediction to longer action-video sequences is also a promising direction.

5 Conclusion

We presented SWAM, a unified observation–action generation framework for visual navigation that jointly predicts and synthesizes intermediate visual observations and corresponding action trajectories. By coupling action and perception in a single-pass inference, SWAM addresses the limitations of candidate-based, verification-centric pipelines, ensuring goal alignment, spatial consistency, and temporal coherence. Our approach leverages spatial priors while requiring only monocular RGB input at inference time. We further propose a visual-guided action refinement module and trajectory-scale regularization loss for improving the action and trajectory prediction performance. Extensive evaluations demonstrate that SWAM surpasses state-of-the-art two-stage baselines in trajectory accuracy, goal success rate, visual fidelity, and inference efficiency, while maintaining stable performance across long-distance and cross-dataset scenarios. These results demonstrate the potential of jointly modeling actions and observations for embodied visual navigation.

Acknowledgements

This work was conducted during an internship at Xiaomi EV. We express our sincere gratitude to the entire team at Xiaomi EV for their generous support and collaboration. This work was supported by the National Natural Science Foundation of China under Grant 62371201.

References

1. Allen, J.F., Koomen, J.A.: Planning using a temporal world model. In: Proceedings of the Eighth international joint conference on Artificial intelligence-Volume 2. pp. 741–747 (1983)
2. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15791–15801 (2025)
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
4. Bonin-Font, F., Ortiz, A., Oliver, G.: Visual navigation for mobile robots: A survey. Journal of intelligent and robotic systems **53**(3), 263–296 (2008)
5. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)
6. Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., Tedrake, R.: Diffusion policy: Visuomotor policy learning via action diffusion. In: Robotics: Science and Systems (RSS) (2023)
7. Ebert, F., Finn, C., Dasari, S., Xie, A., Lee, A., Levine, S.: Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. arXiv preprint arXiv:1812.00568 (2018)
8. Finn, C., Levine, S.: Deep visual foresight for planning robot motion. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 2786–2793. IEEE (2017)
9. Gao, S., Liang, W., Zheng, K., Malik, A., Ye, S., Yu, S., Tseng, W.C., Dong, Y., Mo, K., Lin, C.H., et al.: Dreamdojo: A generalist robot world model from large-scale human videos. arXiv preprint arXiv:2602.06949 (2026)
10. Guo, Y., Hu, Y., Zhang, J., Wang, Y.J., Chen, X., Lu, C., Chen, J.: Prediction with action: Visual policy learning via joint denoising process. Advances in Neural Information Processing Systems **37**, 112386–112410 (2024)
11. Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2616–2625 (2017)
12. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3), 440 (2018)
13. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: International Conference on Learning Representations (ICLR) (2020)
14. Hao, S., Gu, Y., Ma, H., Hong, J., Wang, Z., Wang, D., Hu, Z.: Reasoning with language model is planning with world model. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 8154–8173 (2023)
15. Hirose, N., Shah, D., Sridhar, A., Levine, S.: Sacson: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters **9**(1), 49–56 (2023)
16. Hirose, N., Xia, F., Martín-Martín, R., Sadeghian, A., Savarese, S.: Deep visual mpc-policy learning for navigation. IEEE Robotics and Automation Letters **4**(4), 3184–3191 (2019)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

18. 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)
19. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
20. 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)
21. 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)
22. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023)
23. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
24. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., et al.: Learning to navigate in complex environments. In: International Conference on Learning Representations (2017)
25. Parisotto, E., Salakhutdinov, R.: Neural map: Structured memory for deep reinforcement learning. In: International Conference on Learning Representations (2018)
26. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
27. Savinov, N., Dosovitskiy, A., Koltun, V.: Semi-parametric topological memory for navigation. In: International Conference on Learning Representations (2018)
28. Shah, D., Equi, M.R., Osiński, B., Xia, F., Ichter, B., Levine, S.: Navigation with large language models: Semantic guesswork as a heuristic for planning. In: Conference on Robot Learning. pp. 2683–2699. PMLR (2023)
29. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Ving: Learning open-world navigation with visual goals. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13215–13222. IEEE (2021)
30. Shah, D., Eysenbach, B., Rhinehart, N., Levine, S.: Rapid exploration for open-world navigation with latent goal models. In: Conference on Robot Learning. pp. 674–684. PMLR (2022)
31. Shah, D., Sridhar, A., Bhorkar, A., Hirose, N., Levine, S.: Gnm: A general navigation model to drive any robot. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7226–7233. IEEE (2023)
32. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: Vint: A foundation model for visual navigation. In: Conference on Robot Learning. pp. 711–733. PMLR (2023)
33. Shen, Y., Wei, F., Du, Z., Liang, Y., Lu, Y., Yang, J., Zheng, N., Guo, B.: Videovla: Video generators can be generalizable robot manipulators. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems(NeurIPS2025) (2025)
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)

35. 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 (2021)
36. Sridhar, A., Shah, D., Glossop, C., Levine, S.: Nomad: Goal masked diffusion policies for navigation and exploration. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 63–70. IEEE (2024)
37. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
38. Triest, S., Sivaprakasam, M., Wang, S.J., Wang, W., Johnson, A.M., Scherer, S.: Tartandrive: A large-scale dataset for learning off-road dynamics models. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2546–2552. IEEE (2022)
39. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
40. Wang, H., Liu, D., Xie, H., Liu, H., Ma, E., Yu, K., Wang, L., Wang, B.: Mila: Multi-view intensive-fidelity long-term video generation world model for autonomous driving. arXiv preprint arXiv:2503.15875 (2025)
41. Wu, Y., Karunratanakul, K., Luo, Z., Tang, S.: Uniphys: Unified planner and controller with diffusion for flexible physics-based character control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13214–13224 (2025)
42. Xiong, Z., Ye, X., Yaman, B., Cheng, S., Lu, Y., Luo, J., Jacobs, N., Ren, L.: UniDrive-WM: Unified understanding, planning and generation world model for autonomous driving. arXiv preprint arXiv:2601.04453 (2026)
43. Yang, J., Chitta, K., Gao, S., Chen, L., Shao, Y., Jia, X., Li, H., Geiger, A., Yue, X., Chen, L.: Resim: Reliable world simulation for autonomous driving. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 167710–167741 (2025)
44. Yang, L., Bai, Y., Eskandar, G., Shen, F., Altillawi, M., Chen, D., Liu, Z., Valada, A.: Covar: Co-generation of video and action for robotic manipulation via multi-modal diffusion. arXiv preprint arXiv:2512.16023 (2025)
45. Yang, Y., Liu, J., Zhang, Z., Zhou, S., Tan, R., Yang, J., Du, Y., Gan, C.: Mindjourney: Test-time scaling with world models for spatial reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
46. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025)
47. Ye, S., Ge, Y., Zheng, K., Gao, S., Yu, S., Kurian, G., Indupuru, S., Tan, Y.L., Zhu, C., Xiang, J., et al.: World action models are zero-shot policies. arXiv preprint arXiv:2602.15922 (2026)
48. Yin, T., Mei, Z., Sun, T., Zha, L., Zhou, E., Bao, J., Yamane, M., Sho, O., Majumdar, A.: Womap: World models for embodied open-vocabulary object localization. In: RSS 2025 Workshop: Mobile Manipulation: Emerging Opportunities & Contemporary Challenges (2025)

49. Zhang, J., Jiang, M., Dai, N., Lu, T., Uzunoglu, A., Zhang, S., Wei, Y., Wang, J., Patel, V.M., Liang, P.P., et al.: World-in-world: World models in a closed-loop world. arXiv preprint arXiv:2510.18135 (2025)
50. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27220–27230 (2025)
51. Zhang, T., Hu, X., Xiao, J., Zhang, G.: A survey of visual navigation: From geometry to embodied ai. Engineering Applications of Artificial Intelligence **114**, 105036 (2022)
52. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
53. Zhou, G., Pan, H., LeCun, Y., Pinto, L.: Dino-wm: World models on pre-trained visual features enable zero-shot planning. arXiv preprint arXiv:2411.04983 (2024)
54. Zhou, S., Du, Y., Yang, Y., Han, L., Chen, P., Yeung, D.Y., Gan, C.: Learning 3d persistent embodied world models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
55. Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., He, T.: Aether: Geometric-aware unified world modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8535–8546 (2025)
56. 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. IEEE (2017)