Aerial World Model for Long-horizon Visual Generation and Navigation in 3D Space

Weichen Zhang*
SIGS, Tsinghua University
Shenzhen, China

Peizhi Tang*
Nanyang Technological University
Singapore, Singapore

Xin Zeng
SIGS, Tsinghua University
Shenzhen, China

Fanhang Man
SIGS, Tsinghua University
Shenzhen, China

Shiquan Yu
SIGS, Tsinghua University
Shenzhen, China

Zichao Dai
University of the Chinese Academy of Sciences
Beijing, China

Baining Zhao
SIGS, Tsinghua University
Shenzhen, China

Wei Wu
Tsinghua University
Beijing, China

Chen Gao†
BNRist, Tsinghua University
Beijing, China

Zhibo Chen
University of Science and Technology of China
Hefei, China

Xin Wang
DCST, BNRist, Tsinghua University
Beijing, China

Xinlei Chen†
SIGS, Tsinghua University
Shenzhen, China

Yong Li
EE, BNRist, Tsinghua University
Beijing, China

Wenwu Zhu
DCST, BNRist, Tsinghua University
Beijing, China

Abstract

Unmanned aerial vehicles (UAVs) have emerged as powerful embodied agents. One of the core abilities is autonomous navigation in large-scale three-dimensional environments. Existing navigation policies, however, are typically optimized for low-level objectives such as obstacle avoidance and trajectory smoothness, lacking the ability to incorporate high-level semantics into planning. To bridge this gap, we propose ANWM, an aerial navigation world model that predicts future visual observations conditioned on past frames and actions, thereby enabling agents to rank candidate trajectories by their semantic plausibility and navigational utility. ANWM is trained on 4-DoF UAV trajectories and introduces a physics-inspired module: Future Frame Projection (FFP), which projects past frames into future viewpoints to provide coarse geometric priors. This module mitigates representational uncertainty in long-distance visual generation and captures the mapping between 3D trajectories and egocentric observations. Empirical results demonstrate that ANWM significantly outperforms existing world models in long-distance visual forecasting and improves UAV navigation success rates in large-scale environments.

CCS Concepts

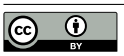
• **Computing methodologies** → *Artificial intelligence.*

Keywords

Aerial Navigation, World Model, Physics-inspired

ACM Reference Format:
Weichen Zhang, Peizhi Tang, Xin Zeng, Fanhang Man, Shiquan Yu, Zichao Dai, Baining Zhao, Wei Wu, Chen Gao, Zhibo Chen, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. 2026. Aerial World Model for Long-horizon Visual Generation and Navigation in 3D Space. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3832540

*Equal contribution
†Corresponding authors: Chen Gao (chgao96@gmail.com), Xinlei Chen (chen.xinlei@sz.tsinghua.edu.cn)



MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3832540

1 Introduction

Unmanned aerial vehicles (UAVs) have emerged as important agents for spatial intelligence with broad applications [43]. A fundamental capability is visual navigation in 3D environments, where UAVs plan paths to efficiently locate visual targets for applications such as object search [13, 51], surveillance [1, 29], and logistics [5, 6].

Early approaches rely on hand-crafted navigation policies [30, 31], which mainly optimize low-level objectives such as obstacle avoidance and smoothness but lack high-level semantic understanding for planning [27, 46]. Inspired by human ability to imagine future scenarios without physical execution [28], recent works [24, 36, 38] leverage world models to predict future visual observations conditioned on trajectories, enabling semantic-aware path planning. However, existing methods [49, 52] are still limited to short-horizon prediction in 2D spaces. For example, NWM [3] generates observations only within a 3-meter range, while Genie 3 [2] supports long-horizon generation but remains restricted to planar actions.

In that case, constructing a world model for visual navigation in aerial spaces has two main challenges. 1) **Complex action space.** Compared to ground robots with only three degrees of freedom (DoF), UAVs have six DoF. Even without considering pitch and roll, the UAV action space remains four-dimensional. Building a world

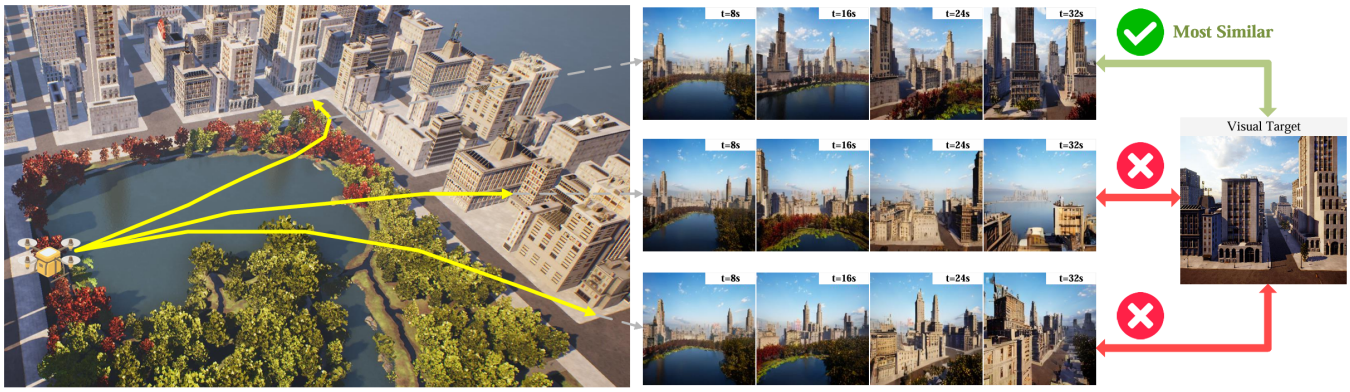


Figure 1: Visual navigation in large-scale aerial space. Given a visual target, the agent is required to plan a trajectory whose final observation aligns with the target. We leverage a world model that imagines visual observations along all possible trajectories. By computing the similarity between the imagined observations and the target, the optimal trajectory is determined. This imagination-based planning paradigm potentially reduces the navigation cost in large-scale open 3D environments.

model that can accurately map such a high-dimensional action space to corresponding visual observations is inherently difficult. 2) **Long-horizon visual generation.** Unlike indoor navigation, aerial navigation typically involves long-horizon locomotion, where the visual target is usually beyond the current field of view and often over 100 meters away from the UAV. Therefore, long horizon refers not only to the temporal dimension but also to the spatial extent. Ensuring long-horizon spatial and temporal consistency in generated visual observations is particularly challenging.

To address the challenges above, we propose an Aerial Navigation World Model (ANWM) that predicts future visual observations conditioned on past observations and future trajectories. To tackle the complex 3D action space, we introduce a 3D visual navigation benchmark that enables ANWM to learn the mapping from 3D actions to aerial observations. To ensure long-horizon spatial–temporal consistency, we design a Future Frame Projection (FFP) module that projects past frames into future perspectives, enforcing visual consistency between the generated future observations and historical ones within their overlapping field of view. Once trained, ANWM is used to predict visual observations along candidate trajectories generated by the path planning policy, enabling the agent to rank trajectories most likely to reach the target.

ANWM is conceptually related to recent diffusion-based world models for navigation and interactive tasks, such as NWM [3] and Matrix-Game [52]. However, unlike these approaches, ANWM is specifically trained to generate first-person visual observations of aerial agents operating in large-scale 3D environments, which introduces unique challenges as discussed above. The main contributions of this paper are as follows:

- We introduce a large-scale dataset for aerial visual generation and navigation, containing 350k trajectory segments with corresponding visual observations.
- The first action-conditioned world model in aerial space, capable of predicting long-horizon visual observations from 3D actions.
- Experimental results demonstrate that our proposed ANWM exhibits spatio-temporally consistent capability in long-horizon navigation, outperforming existing action-conditioned world models.

2 Related Work

Interactive World Models. Recent advances in world modeling have sought to endow agents with a unified representation of perception, action, and prediction within dynamic environments. Early generative approaches [11, 19, 25, 35, 44] such as iVideoGPT [41], MineWorld [15] demonstrated that large video transformers can implicitly capture physical dynamics and object interactions from raw visual sequences, laying the foundation for learning predictive simulators [7, 22, 47, 48] from pixels. Building upon this, GAIA-1 [17] and GAIA-2 [26] introduced large-scale multimodal world models capable of performing tool-augmented reasoning and web-based information synthesis, moving beyond static simulation toward *interactive reasoning*. These models highlight a transition from passive next-frame prediction to active inference—where the agent continuously updates its internal world model based on feedback, search results, or user-provided context. Recent works [49, 52] such as YUME [24], and Genie 3 [2] further extend this paradigm by incorporating generative imagination and high-fidelity visual synthesis, enabling real-time interaction and controllable environment simulation that bridges the gap between embodied intelligence and creative reasoning.

World Models for Navigation. The core insight of leveraging world models for navigation lies in generating observations of unobserved scenes, enabling the agent to perform look-ahead planning for future trajectories. PathDreamer [18] first introduced this idea by generating panoramic observations for indoor waypoint prediction. Follow-up methods such as DreamWalker [36], DreamNav [38], and UniWM [9] further enhanced long-term planning through future-scene imagination. To reduce the complexity of pixel-level generation, NavMorph [45] predicts future latent world states instead of images, while HNR [40] employs NeRF-based latent representations for efficient semantic encoding. Besides, Airscape [53] uses textual instruction to generate aerial observation for further navigation. Beyond next-step prediction, NWM [3] and DreamNav [38] generate global future observations conditioned on entire candidate trajectories, supporting global planning. In parallel, panoramic world models like PanoGen [20] and WCGEN [54]

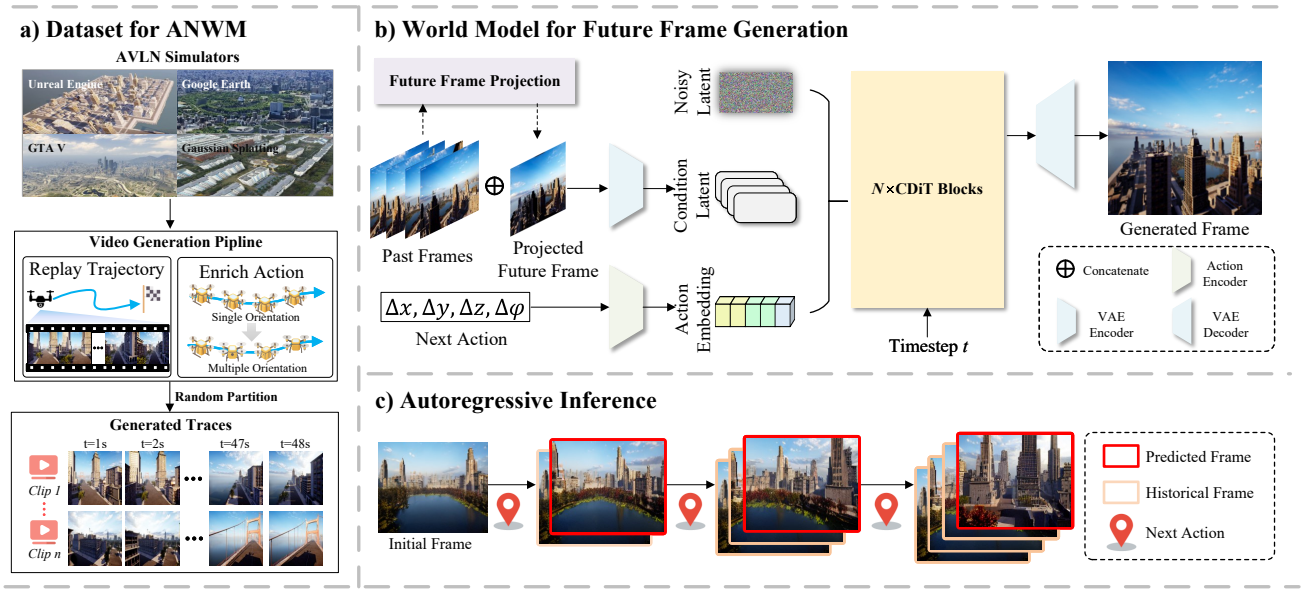


Figure 2: Framework overview. a) Trajectory clips are constructed from AVLN simulators via action enrichment and random partition. b) ANWM generates future observations conditioned on noisy latent, history, projected future frames, and action embeddings, where Future Frame Projection provides scene priors by viewpoint warping. c) Long-horizon generation is performed autoregressively with an updated historical frame queue.

synthesize text-conditioned indoor environments to mitigate data scarcity in VLN benchmarks.

Despite these advances, existing approaches remain largely confined to indoor, 2D settings. Extending these world-model-based imagination and navigation frameworks to outdoor, large-scale, and 3D open spaces remains a significant open challenge.

3 ANWM: Aerial Navigation World Model

3.1 Formulation

In this section, we describe the formulation of visual navigation settings and ANWM.

In the visual navigation task, an agent is required to search for a target specified by an image. The objective is to navigate to a position where the agent's visual observation most closely resembles the target image. Note that the target may or may not be visible from the agent's current viewpoint, which distinguishes this problem from conventional path planning in robotics. In this case, the visual navigation problem can be formally formulated as follows.

Given the agent's current egocentric observation $\mathbf{v}_t \in \mathbb{R}^{H\times W\times 3}$ and its action space $\mathbb{A}$, the agent plans its future actions $\mathbf{D}=(\mathbf{a}_{t+1}, ..., \mathbf{a}_n)$ to reach a location with a final observation $\mathbf{v}_n$ closely resembles the target image $\mathbf{v}^*$. $\mathbf{a}_k \in \mathbb{A}$ is the basic action command of the agent given by relative translation $(\Delta x, \Delta y, \Delta z) \in \mathbb{R}^3$ and yaw change $\Delta\varphi \in \mathbb{R}$. The objective can be formulated as:

$$\begin{aligned} \mathbf{D}^* = (\mathbf{a}^*_{t+1}, \ldots, \mathbf{a}^*_n) = \arg\max_{\mathbf{D}\in\mathbb{D}} \mathcal{S}(\mathbf{v}_n, \mathbf{v}^*) \\ \text{s.t. } \mathbf{v}_n = \mathcal{F}(\mathbf{D}, \mathbf{v}_t) \end{aligned} \tag{1}$$

where $\mathcal{S} : (\mathbf{v}_i, \mathbf{v}_j) \mapsto \mathbb{R}$ is the similarity scores between two latent states, $\mathbb{D}$ is the set of agent's possible trajectories and $\mathcal{F}$ is the agent's kinematic model. Since exhaustively exploring all possible trajectories in open environments would incur prohibitive navigation time and costs, ANWM aims to exploit the counterfactual reasoning capability of generative world models. It enables the agent to *imagine* future observations without executing real actions as depicted in Figure 1. Therefore, the objective of ANWM is to learn a world model W that accurately simulates the distribution of observations conditioned on historical observations and actions:

$$\mathbf{v}_{k+1} \sim W_\theta(\mathbf{v}_{k+1} \mid \mathbf{v}_k, a_{k+1}) \tag{2}$$

where θ denotes the model parameters. We also assume a m-order Markov property in the agent's visual observations, such that the $\mathbf{v}_{k+1}$ depends only on the most recent m observations $\mathbf{v}_{k-m:k}$. Accordingly, Equation 2 becomes:

$$\mathbf{v}_{k+1} \sim W_\theta(\mathbf{v}_{k+1} \mid \mathbf{v}_{k-m:k}, a_{k+1}) \tag{3}$$

Thus, ANWM can generate all observations along the trajectory in an autoregressive manner. By comparing the final generated observation with the target observation, ANWM selects the trajectory with the highest similarity score as the navigation path to be executed.

3.2 Dataset for Aerial Navigation World Model

We first introduce an egocentric aerial agent video dataset, together with well-aligned trajectories, designed for both training and evaluation. As illustrated in Figure 2(a), we first collect UAV trajectories from the aerial vision-and-language navigation (AVLN) benchmarks, including AerialVLN dataset [23], OpenFly dataset [14], and

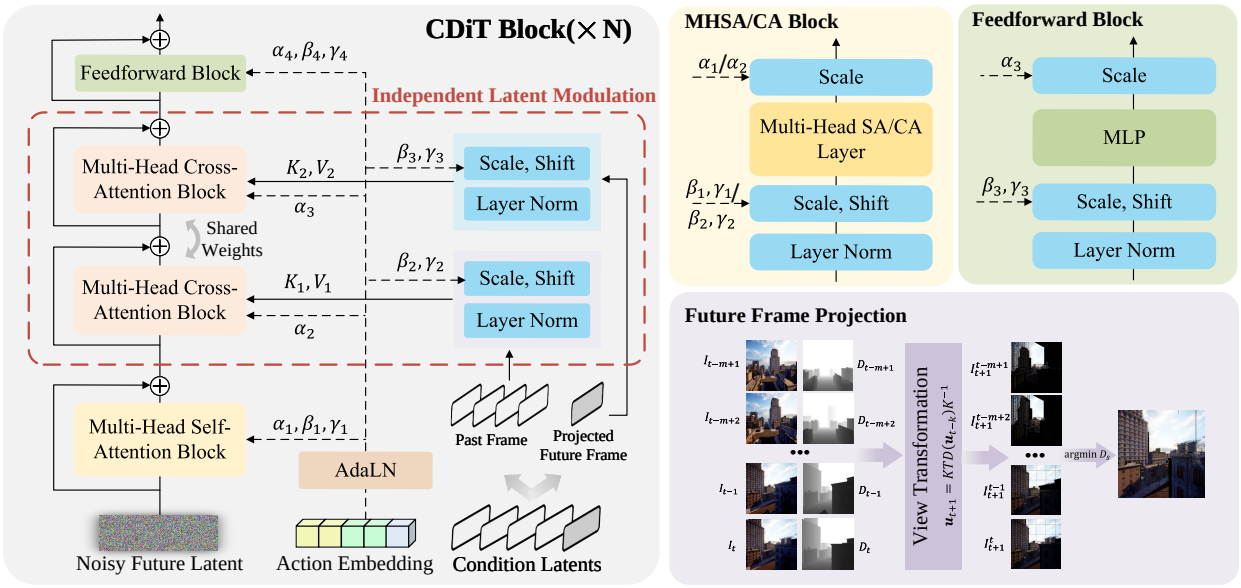


Figure 3: Model Architecture. ANWM adopts CDiT [3] as the backbone but uses the past frame and the projected future frame as distinct conditional signals to control the generation process. Specifically, ANWM first splits the condition latents into the past-frame latent and the projected future-frame latent, and applies separate scale and shift parameters to modulate the strength of the conditioning signal. The modulated latents are then fed into two shared-weight Multi-Head Cross-Attention branches.

OpenUAV dataset [37]. These benchmarks collectively provide over 20k diverse aerial trajectories, spanning more than 40 simulated urban environments constructed using Unreal Engine [10]. Each trajectory is represented as a sequence of 3D waypoints and is paired with a natural language instruction, enabling grounded navigation learning.

To transform these trajectories into video-based training data, we replay the UAV flights within Unreal Engine to capture temporally aligned RGB-D observations from an egocentric perspective. This process ensures that each frame is tightly coupled with the corresponding agent state and action, forming a coherent multimodal sequence suitable for downstream learning.

However, we observe that the original trajectories are heavily biased toward forward movements, which limits the diversity of action distributions and may hinder the model's generalization ability. To address this issue, we propose an action enrichment strategy. Specifically, during trajectory replay, we simultaneously record not only the front view but also additional viewpoints from the left, right, and rear directions. With this strategy, a forward motion in the front view can also be interpreted as a lateral movement in the side views or a backward motion in the rear view. As a result, we obtain a more balanced and expressive dataset without modifying the underlying trajectories.

Finally, we construct a large-scale dataset consisting of 350k trajectory segments for training and 2.2k segments for testing, including 1.1k 2D and 1.1k 3D segments. Each segment comprises a sequence of 47 actions (corresponding to 48 waypoints) selected from a discrete action set, including forward/backward, left/right, up/down translations, and left/right rotations. At each waypoint, we record the corresponding RGB-D observation, resulting in 48 frames per segment. The UAV motion is parameterized with fixed velocities of 5 m/s for horizontal movement, 2 m/s for vertical movement, and 15°/s for rotational motion, leading to an average trajectory length of approximately 80.7 meters. This standardized setup facilitates both stable training and fair evaluation across different models.

3.3 Navigation Framework Overview

World Model for Future Frame Generation. As illustrated in Figure 2, ANWM generates future frames conditioned on historical observations and the next action. A pretrained VAE encoder [4] first compresses frames into latent representations with an 8×8 downsampling factor. The noisy future latent and condition latents are jointly processed by N Conditional Diffusion Transformer (CDiT) blocks and decoded back to pixels by the VAE decoder. The action $a \in \mathbb{R}^4$ is embedded via sine-cosine features and injected into CDiT blocks through adaptive layer normalization. To improve long-horizon generation, we introduce the Future Frame Projection (FFP) module, which provides a coarse future-view prior by projecting historical frames into the next viewpoint. The projected frame is encoded by the VAE and concatenated with historical latents as additional conditioning information, enabling the model to better preserve cross-view consistency.

Autoregressive Inference and Path Planning. After training, we leverage the model to assist path planning for aerial visual navigation as depicted in Figure 2c). We first leverage a heuristic path planner to produce l candidate trajectories by sampling actions from a predefined action set. Then, Gaussian noise is applied as perturbations to the waypoints along each trajectory to

further enhance trajectory diversity. ANWM is leveraged to generate the visual observation at the endpoint of each trajectory to rank these trajectories. During the generation phase, each waypoint pose (x, y, z, φ) is quantized into the relative transformation with respect to the previous waypoint, expressed as $(\Delta x, \Delta y, \Delta z, \Delta\varphi)$, to align with the input format of ANWM. It predicts the next frame in an autoregressive manner, where each newly generated frame is appended to the history buffer, and the most recent m frames are encoded as the state context for subsequent frame generation. Once the final frame of a trajectory is generated, the model evaluates the perceptual similarity $\mathcal{S}(v_n, v^*)$ between the predicted last frame v_n and the target frame v^* using LPIPS. Among all candidate trajectories, the one with the lowest similarity error is then selected as the final navigation path. By default, the number of candidate trajectories l is set to 5.

Future Frame Projection (FFP). Rather than directly feeding the past frame latents into the world model, we design the FFP module that generates a coarse future frame prior, which is then concatenated with the past frame latents as the conditional input. This module leverages the view transformation method in 3D vision that projects m past frames $I_{t-m+1:t}$ into the UAV's future viewpoint at time $t+1$ to obtain an estimated target frame $\tilde{I}_{t+1}$. Given a source frame I_{t-k} with depth D_{t-k}, camera intrinsics K, and relative pose $T_{t-k\to t+1}$, each pixel $\mathbf{u} = [x, y, 1]^\top$ is first back-projected into 3D space via: $\mathbf{p}(x, y) = D_{t-k}(x, y)\, K^{-1}\mathbf{u}$. Then, the 3D point $\mathbf{p}$ is transformed into the target image plane via $\tilde{\mathbf{u}} = KT_{t-k\to t+1}\,\mathbf{p}$, where $\tilde{\mathbf{u}} = [\tilde{x}\tilde{z}, \tilde{y}\tilde{z}, \tilde{z}]$. Using this view transformation function, we project the pixels in I_{t-k} into the target frame: $I_{t+1}^{t-k}(\tilde{u}, \tilde{v}) = I_{t-k}(u, v)$. Each projected target frame I_{t+1}^{t-k} contains only a subset of pixels in the target frame and exhibits substantial missing pixels. To obtain a more complete estimation of the future frame, we fuse all projected frames $I_{t+1}^{t-m+1:t}$ into a single final target frame to compensate for these missing pixels. Specifically, among all projected frames, we select the pixel value corresponding to the minimum depth across frames as the final pixel value of the target frame, which is given by:

$$\begin{aligned} \tilde{I}_{t+1}(x, y) &= I_{t+1}^{t-s^*(x,y)}(x, y), \\ s^*(x, y) &= \arg\min_{s\in[0:m-1]} D_s(x, y). \end{aligned} \tag{4}$$

In that case, we obtain a coarse estimation of the target frame by leveraging the visual cues from all historical observations, which provide an essential prior for future frame generation.

Independent Latent Modulation. After obtaining the estimated future frame $\tilde{I}_{t+1}$, it is encoded with the past frames $I_{t-m:t}$ by the VAE to form the conditional latents $x_{\text{cond}} = \text{VAE}([I_{t-m:t}; \tilde{I}_{t+1}]) = ([x_{t-m:t}; \tilde{x}_{t+1}]) \in \mathbb{R}^{B\times(m+1)\times C\times H\times W}$ which serves as the control signal for future frame generation. We use the Conditional Diffusion Transformer (CDiT) [3] as our aerial world model backbone. As depicted in Figure 3, given an input of noisy future latent $x'_{t+1} \in \mathbb{R}^{B\times C\times H\times W}$, condition latents x_{cond} and an action embedding v_a, CDiT model predicts the denoised future latent x_{t+1} by applying N CDiT blocks over the input latents, where B and C are batch size and channels. In each CDiT block, $v_a \in \mathbb{R}^d$ is used to generate scale $\boldsymbol{\alpha} \in \mathbb{R}^{4\times d_e}$, $\boldsymbol{\beta} \in \mathbb{R}^{5\times d_e}$ and shift $\boldsymbol{\gamma} \in \mathbb{R}^{5\times d_e}$ coefficients by AdaLN [42] block:

$$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} = \text{AdaLN}(\text{SiLU}(v_a)), \tag{5}$$

where d_e is the coefficient dimension. Although $x_{t-m:t}$ and $\tilde{x}_{t+1}$ both represent the agent's observations at different time steps, they exhibit distinct feature distributions. The past frames are real observations and always semantically meaningful, while the projected future frame is a synthesized image with projection errors and can even become meaningless if no overlapping field of view between the two perspectives. Therefore, we propose the Independent Latent Modulation (ILM) method to modulate the distributions of $x_{t-m:t}$ and $\tilde{x}_{t+1}$ separately. Specifically, these latents are passed into two separate modulation layers:

$$\begin{aligned} z_{t-m:t} &= (1+\beta_2)\text{LN}(x_{t-m:t}) + \gamma_2, \\ \tilde{z}_{t+1} &= (1+\beta_3)\text{LN}(\tilde{x}_{t+1}) + \gamma_3, \end{aligned} \tag{6}$$

The modulated condition latents are fed into two shared-weight MHCA blocks sequentially, which is given by:

$$\begin{aligned} z'_{t+1} &= x'_{t+1} + \alpha_1\text{MHSA}((1+\beta_1)\text{LN}(x'_{t+1}) + \gamma_1), \\ z''_{t+1} &= z'_{t+1} + \alpha_2\text{MHCA}(Q_1{=}z'_{t+1}, K_1{=}V_1{=}z_{t-m:t}), \\ z'''_{t+1} &= z''_{t+1} + \alpha_3\text{MHCA}(Q_2{=}z''_{t+1}, K_2{=}V_2{=}\tilde{z}_{t+1}). \end{aligned} \tag{7}$$

Finally, the intermediate latent is processed by a feedforward block to produce the denoised latent z_{t+1}:

$$z_{t+1} = z'''_{t+1} + \alpha_4\text{MLP}((1+\beta_4)\text{LN}(z'''_{t+1}) + \gamma_4). \tag{8}$$

4 Experiments

4.1 Experiment Setup

Dataset and Benchmark We train and evaluate the performance of ANWM on both simulated and real-world datasets. For the simulated setting, we adopt the 1.1k 2D trajectory segments and 1.1k 3D trajectory segments introduced in Section 3. In both the 2D and 3D scenarios, 1,000 segments are used to assess the model's generative capability, while the remaining 100 segments are reserved for evaluating navigation performance, following the experimental protocol of NWM [3]. For real-world evaluation, we utilize Sekai [21], a first-person view video dataset. Specifically, we use its drone-view video subset for both training and testing.

Baselines We compare our method against three representative world models that generate future observations conditioned on action inputs: NWM [3], Matrix-Game [52], and YUME [24]. Since the original architectures of these baselines only support 2D action inputs, we first compare their performance in visual generation and navigation on 2D trajectories. For evaluation on 3D trajectories, we extend the action interface of NWM to accommodate 3D motion and retrain it on our dataset. Retraining Matrix-Game or YUME is not feasible because their source codes are not publicly available. In addition, we compute the average motion velocity of the baseline agents and generate videos of varying durations to ensure that the distance traveled within the same time interval is consistent with that of ANWM.

Metrics For the generation task, we employ FID [16], DreamSim [12], and LPIPS [50] to evaluate the semantic fidelity of the generated results, and use MSE, SSIM, and PSNR to assess their pixel-level accuracy. For the navigation task, we use Absolute Translation

Table 1: Visual generation results of 2D and 3D trajectories.

Timeshift	Method	2D (Simulation)			3D (Simulation)			3D (Real World)		
		LPIPS ↓	DreamSim ↓	FID ↓	LPIPS ↓	DreamSim ↓	FID ↓	LPIPS ↓	DreamSim ↓	FID ↓
4s	Matrix-Game	0.589	0.222	40.2	-	-	-	-	-	-
	YUME	0.571	0.196	73.0	-	-	-	-	-	-
	NWM	0.377	0.259	38.5	0.376	0.247	39.9	0.232	0.183	124.3
	ANWM (ours)	**0.184**	**0.125**	**19.2**	**0.192**	**0.143**	**22.6**	**0.204**	**0.149**	**111.5**
8s	Matrix-Game	0.642	0.307	43.6	-	-	-	-	-	-
	YUME	0.694	0.375	143.1	-	-	-	-	-	-
	NWM	0.422	0.291	43.5	0.428	0.280	38.9	0.250	0.200	131.5
	ANWM (ours)	**0.226**	**0.148**	**20.7**	**0.236**	**0.170**	**25.1**	**0.220**	**0.147**	**117.2**
16s	Matrix-Game	0.714	0.477	71.1	-	-	-	-	-	-
	YUME	0.853	0.701	232.3	-	-	-	-	-	-
	NWM	0.470	0.336	46.3	0.482	0.321	42.1	0.295	0.202	138.2
	ANWM (ours)	**0.313**	**0.202**	**24.5**	**0.301**	**0.210**	**29.4**	**0.256**	**0.158**	**125.0**
32s	Matrix-Game	0.790	0.646	135.9	-	-	-	-	-	-
	YUME	0.902	0.787	269.7	-	-	-	-	-	-
	NWM	0.524	0.400	61.0	0.535	0.377	47.6	0.388	0.229	149.6
	ANWM (ours)	**0.433**	**0.294**	**32.5**	**0.389**	**0.271**	**36.1**	**0.371**	**0.176**	**138.9**

Table 2: 2D and 3D navigation results.

Method	2D (Simulation)				3D (Simulation)				3D (Real World)			
	ATE ↓	RPE ↓	SR ↑	NE ↓	ATE ↓	RPE ↓	SR ↑	NE ↓	ATE ↓	RPE ↓	SR ↑	NE ↓
YUME	15.92	1.80	0.0	24.32	-	-	-	-	-	-	-	-
Matrix-Game	14.75	1.53	16.0	20.98	-	-	-	-	-	-	-	-
NWM	7.72	0.89	63.0	12.71	8.52	**1.03**	58.0	14.51	17.13	3.79	28.0	27.02
ANWM (ours)	**6.30**	**0.78**	**73.0**	**10.30**	**8.13**	1.06	**60.0**	**14.12**	**15.56**	**3.51**	**33.0**	**24.41**

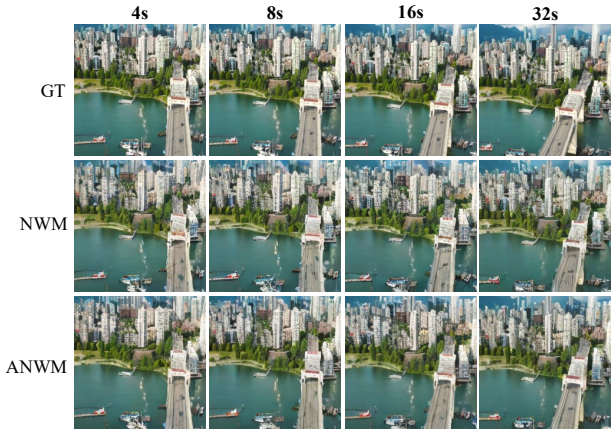


Figure 4: Visual generation in the real-world dataset

Error (ATE), Relative Pose Error (RPE) [33], Success Rate (SR) [23], and Navigation Error (NE) [14] to evaluate navigation accuracy.

Implementation Details. Since the real-world dataset Sekai [21] does not provide depth images, we use Pi-long [8, 39] to estimate the depth information for the image sequences. The input and output frames are resized to a resolution of 224×224. ANWM is implemented with 8 CDiT blocks and trained for 300k steps on four NVIDIA A800 40GB GPUs. We use the AdamW optimizer with a learning rate of $8e-5$. By default, the number of conditional past frames, also referred to as the context size m is 4. During inference, ANWM autoregressively generates the visual observation at each waypoint along the trajectory. For each trajectory segment containing 48 frames, we set the first 16 frames as historical observations, and the model is required to predict the next 32 frames.

4.2 Main Results

Visual Generation We report the generation results at 4s, 8s, 16s, and 32s in Table 1. We observe that the performance of all methods degrades as the trajectory length increases, indicating the challenge of long-horizon observation generation. Our method consistently achieves the best performance on both 2D and 3D trajectories, demonstrating superior action-observation correspondence and long-range consistency. In contrast, YUME and NWM exhibit significant performance drops at 16s and 32s, showing limitations in maintaining visual consistency over long trajectories. These results indicate that propagating historical scene information effectively improves long-horizon generation.

For real-world videos, LPIPS and DreamSim remain comparable to simulation results, while FID increases substantially. We attribute this gap to the higher variability of real-world data, including diverse backgrounds, lighting, and sensor noise, which enlarges the distribution gap between generated and ground-truth videos.

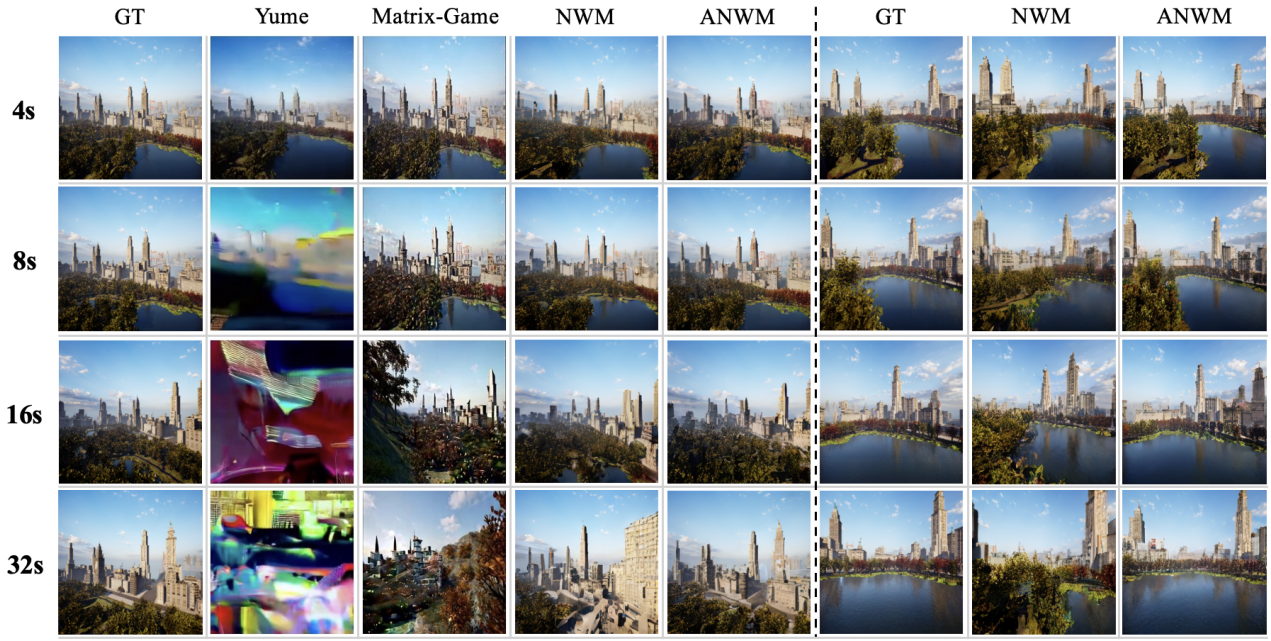


Figure 5: The qualitative results of generated visual observation in the simulated dataset. Left: 2D trajectory. Right: 3D trajectory.

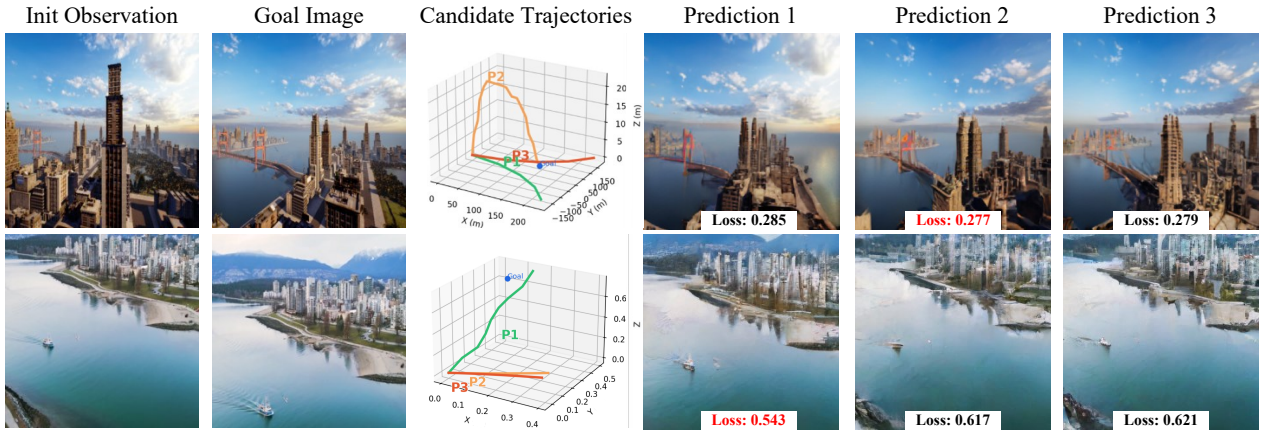


Figure 6: The qualitative results of visual navigation. ANWM ranks each trajectory's final prediction by measuring the LPIPS similarity with the goal image. The trajectory with the lowest LPIPS is selected for execution. The upper and bottom rows are simulated and real-world results, respectively. We only visualize the top-3 trajectories.

We present qualitative results in Figure 4 and Figure 5. The observations generated by our method are more consistent with the ground truth and exhibit higher visual realism across both real-world and simulated datasets. Although NWM and Matrix-Game can produce visually plausible images, their results gradually deviate from the actual motion trajectory as the path length increases. In contrast, YUME suffers from mode collapse at the early stage of generation. Even for 3D trajectories with large altitude variations, ANWM can maintain accurate correspondence between the generated observations and the underlying motion trajectory.

Navigation As depicted in Table 2, ANWM achieves the highest navigation success rate and the lowest navigation error in both 2D and 3D navigation tasks. Specifically, the ATE of ANWM is reduced by 5.1% compared to the second-best method, while the SR is improved by 10% For 3D navigation results, our method outperforms NWM by at least 2% in SR and 4.7% in ATE across both simulated and real-world datasets, further demonstrating the effectiveness and robustness of our approach for long-range navigation.

We provide qualitative navigation results in Figure 6. The heuristic path planner generates five candidate trajectories, and ANWM ranks them based on the LPIPS similarity between each trajectory's final predicted observation and the goal image, selecting the top-ranked trajectory as the final path. Benefiting from its long-horizon visual generation capability, ANWM is able to select the navigation path that is closest to the goal in the 3D environment based on visual similarity, thereby improving the navigation success rate.

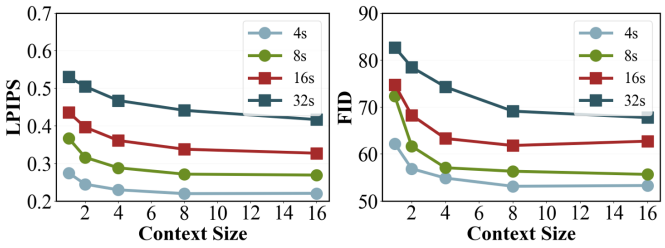


Figure 7: Ablation of context size for future frame projection.

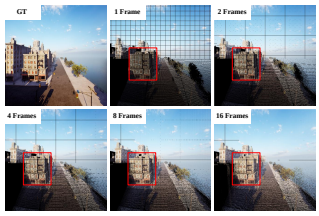


Figure 8: Qualitative results of FFP via different context sizes.

Table 3: Quantitative results of FFP via different context sizes.

Count	MSE	SSIM	PSNR
1	0.287	0.449	5.63
2	0.262	0.464	6.13
4	0.249	0.568	6.36
8	0.236	0.627	6.59
16	0.217	0.651	6.90

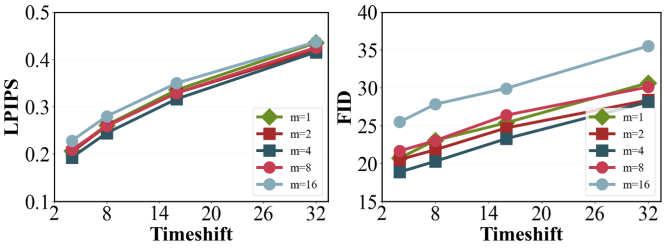


Figure 9: Ablation of context size for generation.

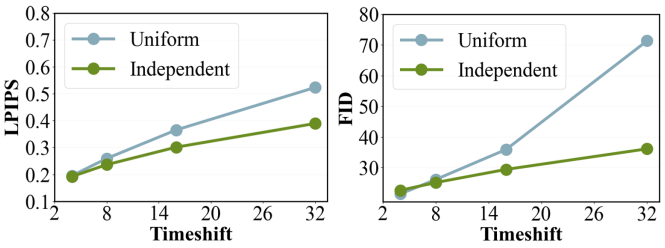


Figure 10: Ablation of condition latents modulation.

4.3 Ablation Study

In this section, we decouple the context size into two factors: (1) the number of historical frames n used for FFP, and (2) the number of historical latent frames m used for denoising.

Context Size for Future Frame Projection We fix $m = 16$ and vary the number of projected past frames n from 1 to 16 for FFP. As shown in Figure 7, increasing n consistently improves generation quality for both short-term (4s) and long-term (32s) predictions. The qualitative results in Figure 8 and metrics in Table 3 show that more historical frames provide richer scene context, producing more accurate future projections and improving generation performance.

Context Size for Generation We fix $n = 1$ and train the model with different context sizes $m = 1, 2, 4, 8, 16$. The results are shown in Figure 9. Although prior work [3] suggests that increasing the context size improves generative performance when the context size is not bigger than 4, our experiments show that the performance degrades when the context size is extended to 16. We assume this is because distant historical frames differ significantly from the future frame, introducing additional noise to the generation process.

Modulation Method for Condition Latents We train the model with both uniform and independent latent modulation architectures. The former modulates the condition latents of past frames and the projected future frame using the same scale and shift parameters, while the latter uses two separate modulation modules. The results shown in Figure 10 indicate that the two methods perform comparably in short-range generation. However, for long-horizon generation, the independent modulation significantly outperforms the uniform modulation.

5 Limitations

In this section, we identify several limitations of our approach. While our method can generate realistic visual observations along trajectories of approximately 100 meters, it tends to experience mode collapse [32, 34] when extended to longer distances (e.g., around 200 meters). We hypothesize that the accumulated viewpoint variations over such long trajectories lead to large discrepancies between future and historical observations, making the past frames ineffective as priors for future frame prediction. Second, our method occasionally produces distortions in the generation of fine-grained texture, such as in the details of the windows or facades of the buildings in Figure 11. To alleviate this issue, we plan to incorporate additional physical constraints to enhance the model's perception of three-dimensional spatial structure. Besides, the world model is currently used to rank different trajectory candidates. In future work, we plan to enable the world model to assist the UAV in actively planning its own paths.

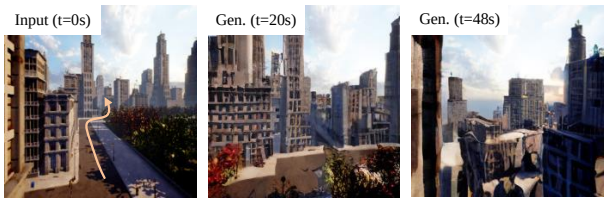


Figure 11: Limitations of our model. ANWM fails to generate fine-grained architectural textures (middle) and consistent observations for extremely long-range trajectories (right).

6 Conclusion

In this work, we present ANWM, the aerial world model capable of generating long-horizon visual observations along 3D trajectories for UAV navigation. Experimental results demonstrate the effectiveness of ANWM in long-range visual generation and 3D navigation accuracy. We also discuss several limitations of ANWM in generating fine-grained textures and providing timely guidance for UAV path planning. We plan to incorporate additional physical constraints and 3D path planning algorithms to address these limitations, which we leave for future work.

Acknowledgments

This paper was supported by the Natural Science Foundation of China under Grant 62371269, Shenzhen Low-Altitude Airspace Strategic Program Portfolio No. Z25306110. This work is also supported in part by Tsinghua University Toyota Research Center.

References

[1] Tazeem Ahmad, Alicia Morel, Nuo Cheng, Kannappan Palaniappan, Prasad Calyam, Kun Sun, and Jianli Pan. 2025. Future UAV/Drone Systems for Intelligent Active Surveillance and Monitoring. *Comput. Surveys* 58, 2 (2025), 1–37.

[2] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. 2025. Genie 3: A New Frontier for World Models. (2025).

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

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[5] Xuecheng Chen, Haoyang Wang, Yuhan Cheng, Haohao Fu, Yuxuan Liu, Fan Dang, Yunhao Liu, Jinqiang Cui, and Xinlei Chen. 2024. Ddl: Empowering delivery drones with large-scale urban sensing capability. *IEEE Journal of Selected Topics in Signal Processing* (2024).

[6] Xuecheng Chen, Haoyang Wang, Zuxin Li, Wenbo Ding, Fan Dang, Chengye Wu, and Xinlei Chen. 2022. Deliversense: Efficient delivery drone scheduling for crowdsensing with deep reinforcement learning. In *Adjunct proceedings of the 2022 ACM international joint conference on pervasive and ubiquitous computing and the 2022 ACM international symposium on wearable computers*. 403–408.

[7] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. 2023. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384* (2023).

[8] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. 2025. VGGT-Long: Chunk it, Loop it, Align it – Pushing VGGT's Limits on Kilometer-scale Long RGB Sequences. arXiv:2507.16443 [cs.CV] https://arxiv.org/abs/2507.16443

[9] Yifei Dong, Fengyi Wu, Guangyu Chen, Zhi-Qi Cheng, Qiyu Hu, Yuxuan Zhou, Jingdong Sun, Jun-Yan He, Qi Dai, and Alexander G Hauptmann. 2025. Unified World Models: Memory-Augmented Planning and Foresight for Visual Navigation. *arXiv preprint arXiv:2510.08713* (2025).

[10] Epic Games. 2019. *Unreal Engine*. https://www.unrealengine.com

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

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

[13] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 2023. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23171–23181.

[14] Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, et al. 2025. OpenFly: A Comprehensive Platform for Aerial Vision-Language Navigation. *arXiv preprint arXiv:2502.18041* (2025).

[15] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. 2025. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388* (2025).

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

[17] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080* (2023).

[18] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. 2021. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14738–14748.

[19] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).

[20] Jialu Li and Mohit Bansal. 2023. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in neural information processing systems* 36 (2023), 21878–21894.

[21] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. 2025. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675* (2025).

[22] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. 2025. Wonderland: Navigating 3d scenes from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 798–810.

[23] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. 2023. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15384–15394.

[24] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. 2025. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744* (2025).

[25] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

[26] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. 2025. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523* (2025).

[27] Cansu Sancaktar, Christian Gumbsch, Andrii Zadaianchuk, Pavel Kolev, and Georg Martius. 2025. Sensei: Semantic exploration guided by foundation models to learn versatile world models. *arXiv preprint arXiv:2503.01584* (2025).

[28] Martin Seeber, Matthias Stangl, Mauricio Vallejo Martelo, Uros Topalovic, Sonja Hiller, Casey H Halpern, Jean-Philippe Langevin, Vikram R Rao, Itzhak Fried, Dawn Eliashiv, et al. 2025. Human neural dynamics of real-world and imagined navigation. *Nature Human Behaviour* 9, 4 (2025), 781–793.

[29] Eduard Semsch, Michal Jakob, Dušan Pavlicek, and Michal Pechoucek. 2009. Autonomous UAV surveillance in complex urban environments. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 2. IEEE, 82–85.

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

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

[32] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. 2017. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems* 30 (2017).

[33] Jürgen Sturm, Wolfram Burgard, and Daniel Cremers. 2012. Evaluating egomotion and structure-from-motion approaches using the TUM RGB-D benchmark. In *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*, Vol. 13. 6.

[34] Hoang Thanh-Tung and Truyen Tran. 2020. Catastrophic forgetting and mode collapse in GANs. In *2020 international joint conference on neural networks (ijcnn)*. IEEE, 1–10.

[35] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).

[36] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. 2023. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10873–10883.

[37] Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. 2024. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087* (2024).

[38] Yunheng Wang, Yuetong Fang, Taowen Wang, Yixiao Feng, Yawen Tan, Shuning Zhang, Peiran Liu, Yiding Ji, and Renjing Xu. 2025. DreamNav: A Trajectory-Based Imaginative Framework for Zero-Shot Vision-and-Language Navigation. *arXiv preprint arXiv:2509.11197* (2025).

[39] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. 2025. π^3: Scalable Permutation-Equivariant Visual Geometry Learning. arXiv:2507.13347 [cs.CV] https://arxiv.org/abs/2507.13347

[40] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, and Shuqiang Jiang. 2024. Lookahead exploration with neural radiance representation for continuous vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13753–13762.

[41] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. 2024. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems* 37 (2024), 68082–68119.

[42] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. Understanding and improving layer normalization. *Advances in neural information processing systems* 32 (2019).

[43] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 10632–10643.

[44] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).

[45] Xuan Yao, Junyu Gao, and Changsheng Xu. 2025. NavMorph: A Self-Evolving World Model for Vision-and-Language Navigation in Continuous Environments. *arXiv preprint arXiv:2506.23468* (2025).

[46] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. 2024. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 42–48.

[47] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. 2025. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5916–5926.

[48] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. 2024. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6658–6667.

[49] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. 2025. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325* (2025).

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[51] Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025. CityNavAgent: Aerial Vision-and-Language Navigation with Hierarchical Semantic Planning and Global Memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 31292–31309. doi:10.18653/v1/2025.acl-long.1511

[52] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. 2025. Matrix-Game: Interactive World Foundation Model. *arXiv preprint arXiv:2506.18701* (2025).

[53] Baining Zhao, Rongze Tang, Mingyuan Jia, Ziyou Wang, Fanhang Man, Xin Zhang, Yu Shang, Weichen Zhang, Wei Wu, Chen Gao, et al. 2025. AirScape: An Aerial Generative World Model with Motion Controllability. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 12519–12528.

[54] Yu Zhong, Rui Zhang, Zihao Zhang, Shuo Wang, Chuan Fang, Xishan Zhang, Jiaming Guo, Shaohui Peng, Di Huang, Yanyang Yan, et al. 2024. World-Consistent Data Generation for Vision-and-Language Navigation. *arXiv preprint arXiv:2412.06413* (2024).