

AirScape: An Aerial Generative World Model with Motion Controllability

Baining Zhao*
Shenzhen International
Graduate School, Tsinghua
University
Pengcheng Laboratory
Shenzhen, China

Rongze Tang*
School of Computer Science
& Technology, Beijing
Institute of Technology
Beijing, China

Mingyuan Jia*
Department of Automation,
Tsinghua University
Beijing, China

Ziyou Wang*
Computer and
Communication Engineering
College, Northeastern
University
Qinhuangdao, China

Fanhang Man
Shenzhen International
Graduate School, Tsinghua
University
Shenzhen, China

Xin Zhang
Manifold AI
Beijing, China

Yu Shang
Department of Electronic
Engineering, Tsinghua
University
Beijing, China

Weichen Zhang
Shenzhen International
Graduate School, Tsinghua
University
Pengcheng Laboratory
Shenzhen, China

Wei Wu
Manifold AI
Beijing, China

Chen Gao[†]
BNRist, Tsinghua University
Beijing, China

Xinlei Chen[†]
Shenzhen International
Graduate School, Tsinghua
University
Shenzhen, China

Yong Li[†]
Department of Electronic
Engineering, Tsinghua
University
BNRist, Tsinghua University
Beijing, China

Abstract

How to enable agents to predict the outcomes of their own motion intentions in three-dimensional space has been a fundamental problem in embodied intelligence. To explore general spatial imagination capability, we present AirScape, the first world model designed for six-degree-of-freedom aerial agents. AirScape predicts future observation sequences based on current visual inputs and motion intentions. Specifically, we construct a dataset for aerial world model training and testing, which consists of 11k video-intention pairs. This dataset includes first-person-view videos capturing diverse drone actions across a wide range of scenarios, with over 1,000 hours spent annotating the corresponding motion intentions. Then we develop a two-phase schedule to train a foundation model—initially devoid of embodied spatial knowledge—into a world model that is controllable by motion intentions and adheres to physical spatio-temporal constraints. Experimental results demonstrate that AirScape significantly outperforms existing foundation models in 3D spatial imagination capabilities, especially with over a 50% improvement in metrics reflecting motion alignment. The project is available at: <https://embodiedcity.github.io/AirScape/>.

*All four authors contributed equally to this research. Rongze Tang and Ziyou Wang conducted this work during their internship at Tsinghua University.

[†]Corresponding authors: Chen Gao (chgao96@gmail.com), Xinlei Chen (chen.xinlei@sz.tsinghua.edu.cn), Yong Li (liyong07@tsinghua.edu.cn)



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '25, Dublin, Ireland*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2035-2/2025/10

<https://doi.org/10.1145/3746027.3758180>

CCS Concepts

• Computing methodologies → Artificial intelligence.

Keywords

Generative World Model; Aerial Space; Motion Controllability

ACM Reference Format:

Baining Zhao, Rongze Tang, Mingyuan Jia, Ziyou Wang, Fanhang Man, Xin Zhang, Yu Shang, Weichen Zhang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. 2025. AirScape: An Aerial Generative World Model with Motion Controllability. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746027.3758180>

1 Introduction

The unprecedented advancements in generative models [9] have catalyzed a paradigm shift in the development of world models, enabling generation and simulation of the real-world environment, with inputs of texts and actions. The simulation reflects a kind of high-level capability, counterfactual reasoning, which enables the simulation and prediction of possible outcomes based on hypothetical conditions or decisions [17]. By comparing results under different assumptions, world models support better decision-making in unknown or complex environments, which is particularly important for downstream applications such as embodied robotics [39, 64], autonomous driving [19, 20], etc. Moreover, the world model can enhance the agents' spatial intelligence [24, 69], with the human-like critical ability in understanding the environment. Specifically, when an agent operates in a three-dimensional real-world space, given its current observation, we expect it to predict how its first-person

Motion Intention: The drone flies to the left while rotating to the right, rotating clockwise about 45 degrees around the pagoda, while keeping the pagoda and surrounding structures centered in the field of view.



Motion Intention: The drone hovers in place while gradually rotating to the left, ends up facing a broader view of the buildings and the street below.



Motion Intention: The drone moves forward while capturing overhead footage of vehicles traveling forward along a bridge, as the camera gimbal angle remains consistent and focused downward to track the white car throughout.



Figure 1: In 3D space, AirScape can predict the sequence of observations that would result if a six-degree-of-freedom aerial agent executed a series of actions to achieve an intention, based on current visual observations. AirScape can handle diverse actions (translation, rotation, and their combinations), environments (rural, urban), viewpoints (top-down, horizon), and lighting conditions (daytime, dusk, nighttime), simulating embodied observation characteristics such as perspective and parallax.

perspective will change after performing actions or tasks, which implicitly indicates how its spatial relationship with the surrounding environment will evolve. This goes beyond basic spatial perception and understanding [13, 73], enabling navigation [3, 71] and task planning [47] in complex and unknown real-world scenarios.

Current research on spatial world models primarily focuses on humanoid robots and autonomous driving applications [23, 51]. However, the world models for humanoid robots emphasize manipulation and indoor environment modeling, while those for autonomous driving focus on predicting driving behaviors and modeling road dynamics. Both of them operate mostly in two-dimensional planes with limited action spaces. With the development of low-altitude economy, the increasing intelligence of aerial agents, such as drones, drive their widespread applications, such as delivery [11, 44], emergency disaster relief [12, 52, 65], and urban pollution management [43, 49, 75]. The research on aerial world models remains unexplored. Moreover, the spatial geometric complexity of embodied spatial counterfactual reasoning in 3D real-world environments with six degrees of freedom (6DoF) is significantly higher, representing a more general type of world model. Examples of the aerial world model are presented in Figure 1.

Vision is one of the fundamental perceptual modalities, and the latest visual observations inherently contain spatial information [72]. Compared to flight control variables or specific trajectory coordinates, we argue that expressing action intentions in textual form aligns more closely with human reasoning processes and offers greater flexibility. Text-based instructions can represent high-level navigation instructions, such as "move to the boat ahead" or "follow the car in front," as well as low-level commands, such as "rotate 90

degrees to the left." When a series of actions is executed in the current spatial context to fulfill an intention, the most direct outcome is a sequence of visual observations. This input-output structure aligns with the framework of video generation models conditioned on both graphical and textual inputs. Recently, video generation foundation models, represented by diffusion [5, 28] and autoregressive models [10, 68], have rapidly advanced and are becoming important tools for implementing world models [1, 38]. Inspired by the scaling law, large foundation models exhibit generalization capabilities [36, 48]. Video generation models can model dynamic changes in temporal sequences, directly simulating visual information of the environment. This capability aligns closely with the requirements of world models for modeling and predicting future spatio-temporal states. However, constructing a generative aerial world model still faces the following challenges:

- **Lack of aerial datasets:** Training world models requires first-person perspective videos and corresponding textual prompts about aerial agents' actions or tasks. Existing datasets are either third-person views or ground-based perspectives from robots or vehicles [8, 16, 31].
- **Distribution gap between video foundation models and world models:** In terms of text input, existing open-source foundation models focus on generating videos from detailed textual descriptions [5, 29], whereas world models rely on concise instructions or action intents. In terms of video, training data for open-source foundation models mostly consists of third-person videos with limited visual changes [35, 36, 66], while embodied first-person perspectives typically have narrower fields of view and larger visual changes, increasing training difficulty.

- **Diversity in generation:** Drones operate in 6DoF with high flexibility [58]. Compared to ground vehicles, generated scenes include lateral translation, in-place rotation, camera gimbal adjustments, and combinations of multiple actions, making generation more challenging. The aerial spatial world model is required to simulate more complex changes in relative position, perspective variation, and parallax effects.

To address these issues, we first introduce an 11k dataset for training aerial world models. We collect videos from three public drone datasets, segment and filter them, and annotate each video clip with its corresponding motion intents using large multimodal models (LMMs) and human refinement (Section 3). Subsequently, we develop a two-stage training schedule: fine-tuning video generation foundation models to adapt to the text and video distributions; rejection sampling and self-play training are employed to further improve generation outputs that violate spatial physical constraints. (Section 4). Experimental results demonstrate that the proposed spatial world model can predict observations in embodied perspectives when performing various actions or tasks (Section 5). The main contributions of this paper are as follows:

- **The first dataset for training and testing generative world models in aerial spaces**, containing 11k video clips with corresponding textual motion intentions.
- **The first generative world model in aerial spaces**, capable of predicting visual observations from controllable motion intentions in three-dimensional spaces.
- Experimental analysis demonstrates that our proposed AirScape exhibits embodied motion-following simulation and prediction capabilities in aerial space scenarios, outperforming existing general video generation models and world models.

2 Related Work

World Model. World models present a grand vision, serving as simulators to support offline training and interaction for agents, while also enabling high-level reasoning and generalization in real-time decision-making [1, 25, 38]. Current research on world models can be categorized into several key areas. First, general world models aim to develop scalable and generalizable representations to simulate and understand complex environments [7, 55, 63]. Second, world models for embodied AI focus on enabling robots to learn and construct world models through interaction with their surroundings, improving manipulation and navigation capabilities [22, 76]. Third, applications in autonomous driving utilize world models to simulate traffic scenarios and enhance driving safety [20, 50, 62]. However, existing world model research has not yet focused on aerial agents [18, 59]. Actually, the aerial agents have the potential to exhibit more generalized spatial intelligence due to their six degrees of freedom in three-dimensional space.

Video Generation. Exemplified by Sora [46], video generation models have garnered significant attention for their highly realistic and lifelike video generation capabilities [15]. Compared to GAN-based approaches [14], diffusion-based models have demonstrated superior performance in generating high-fidelity videos [5, 28, 30]. Additionally, inspired by the success of transformer architectures in large language models (LLMs), several works have explored autoregressive-based approaches for video generation [10, 42, 68],

leveraging their sequential modeling capabilities. In terms of applications, video generation has expanded into diverse directions. Text-to-video generation has made significant strides, with works like Imagination setting new benchmarks for producing high-quality videos from textual prompts [53, 57]. Image-to-video approaches focus on animating static images based on motion descriptions or physical constraints, enabling dynamic visualizations from static inputs [33, 74]. Despite these advancements, current video generation models are primarily designed to visualize input content, relying heavily on detailed descriptions to control video outputs. They often lack the predictive and reasoning capabilities inherent to world models, which are essential for understanding and simulating the consequences of actions, particularly in complex environments like aerial spaces with six degrees of freedom.

3 Dataset for Aerial World Model

We present an 11k embodied aerial agent video dataset along with corresponding annotations of motion intention, aligning the inputs and outputs of the aerial world model. Below, we detail the dataset construction pipeline and dataset statistics.

3.1 Dataset Construction Pipeline

We first gather the egocentric perspective videos shot by the UAVs from open-sourced dataset: UrbanVideo-Bench [72], NAT2021 [67], and WebUAV-3M [70]. These datasets are derived from various tasks, including vision-language navigation, tracking, etc., featuring diverse drone actions. The scenes span over 10 types, such as industrial areas, residential zones, suburbs, and coastal regions. Additionally, they include various weather conditions, such as sunny days and nighttime. We segmented the videos into 129-frame clips and filtered out those that were static or exhibited abrupt changes. Examples are presented in Figure 2.

Subsequently, we employed a vision-language model to infer the drone’s motion intentions. These intentions could range from simple individual actions (e.g., rotating 45 degrees to the left) to specific tasks (e.g., tracking the white car ahead). We designed a straightforward chain-of-thought process, first identifying the action, then summarizing its stopping conditions, and finally merging them into a coherent and logically structured intention prompt.

Finally, we conducted over 1,000 hours of human refinement. Even the most advanced VLMs struggle to accurately infer the agent’s motion from changes in the embodied perspective. Therefore, we focused on correcting the following aspects: incorrect actions, ambiguous descriptions, and imprecise tasks within the descriptions. The pipeline is shown in Figure 3a.

3.2 Dataset Statistics

The statistical properties of the dataset are shown in Figure 3b-d. The dataset’s motion types are categorized into translation, rotation, and compound, while its scenes span 8 major categories, including roadside, suburbs, and riverside. The motion intention prompt lengths follow a near-normal distribution, with a mean of 163.9 and median of 160 characters. Based on the video content, text length, and word cloud analysis, the dataset demonstrates diversity and is well-suited for training and testing models to predict future sequence observations.

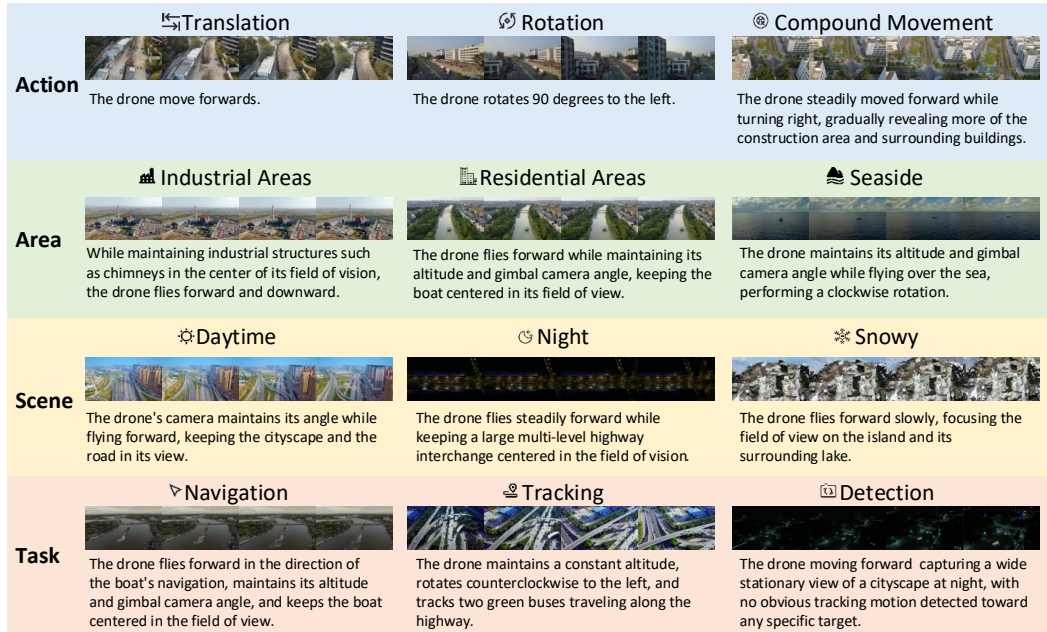


Figure 2: The proposed dataset includes samples with diverse actions, areas, scenes, and tasks.

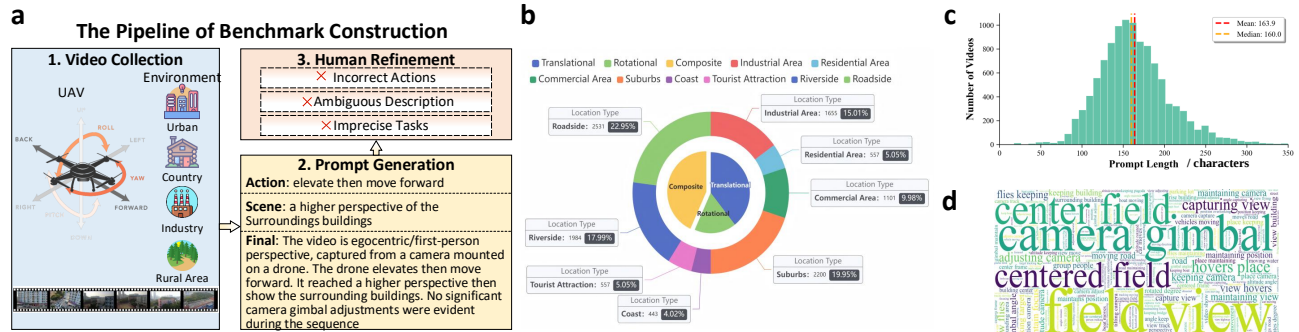


Figure 3: a. Dataset construction pipeline. b. Proportions of different actions and various scenarios in the dataset. c. Length distribution of intention prompts in the dataset. d. Word cloud of intention prompts in the dataset.

4 Learning an Aerial World Model

An Aerial world model W receives the current state of the world and predicts the future state if an aerial agent performs motion intention. In this work, the current state refers to the current egocentric visual observation o . The intention refers to the trajectory, movement, or goal of the aerial agent in 6DoF spaces, expressed in high-level textual form p . The future state represents the sequential changes in embodied visual observations, expressed in the form of a video $\hat{v}$. Thus the above process can be expressed as:

$$\hat{v} = W(o, p) \quad (1)$$

We propose a two-phase training schedule to obtain W , as shown in Figure 4. First, the foundation model is fine-tuned on the proposed dataset to acquire basic intention controllability. Furthermore, a self-play approach is introduced, where synthetic data is generated and trained based on a spatio-temporal discriminator, ensuring the generated videos adhere to spatio-temporal constraints.

4.1 Phase 1: Learning Intention Controllability

Developing a world model based on a pre-trained video generation foundation model can leverage its inherent capability for dynamic modeling in temporal sequences, significantly reducing data and training resource requirements. To enable the video generation foundation model to predict future sequential observations based on the current observation and embodied motion intention, we first perform supervised fine-tuning (SFT). Currently, the foundation model is accustomed to inputs in the form of textual prompts that provide detailed descriptions of the content and specifics of the video to be generated. In this case, the model primarily serves as a tool for visualizing text and images into videos, rather than functioning as the predictive and reasoning world model we aim to develop. For example:

As the perspective moves forward, the blue building ahead gradually enlarges in the field of view. In front of it is a road with a continuous stream of cars ...

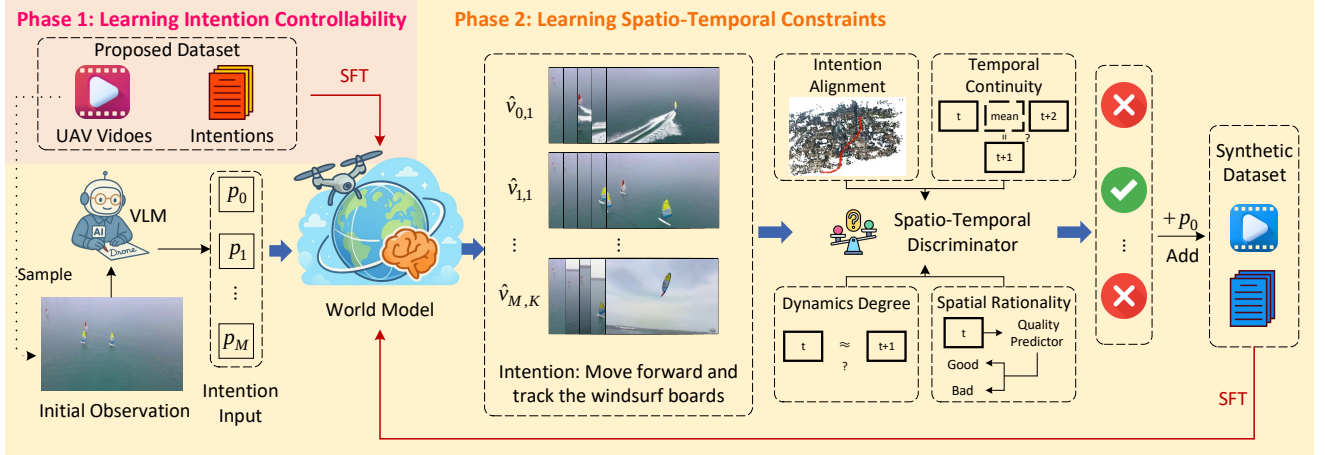


Figure 4: The proposed two-phase training schedule aims to develop an aerial world model that is motion-controllable while adhering to physical spatio-temporal constraints. Phase 1 involves supervised fine-tuning (SFT) on the aerial video-intention pair dataset introduced in Section 3. Phase 2 uses rejection sampling to roll out high-quality samples for iterative SFT. We give an example of this process: The initial frame depicts windsurf boards on the sea, with the drone intending to move forward while keeping them in focus. Among the generated videos, the first is unrealistic as a windsurf board moves like a speedboat, and the last is unreasonable as a board flies into the air. The second video is consistent with real-world physics, with the drone adjusting its gimbal downward to keep the boards in view, making them appear larger in the egocentric perspective.

In contrast, a world model should fully extract the environmental information embedded in the current observation and predict the sequence of observations that would result from an aerial agent executing a series of actions to fulfill its motion intention. For example:

The drone moves forward until it approaches the blue building.

Additionally, the motion of a 6DoF drone involves actions such as lateral translation, vertical movement, rotational adjustments, and gimbal angle changes—scenarios that are underrepresented in the foundation model’s pretraining phase. The combination of these actions further increases the search space for future states.

To empower foundation model with the aerial spatial prediction capability, we perform SFT training on the video generation foundation model using the proposed dataset of video and textual intention pairs $\mathcal{D} = \{(v_i, p_i)\}_{i=1}^N$, where v_i represents a video and p_i is its corresponding textual intention. The fine-tuning process minimizes the reconstruction loss between the predictive outcome $\hat{v}$ and the ground truth outcome v :

$$\mathcal{L}_{\text{fine-tune}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{recon}}(W(o_i, p_i), v_i), \quad (2)$$

where $\mathcal{L}_{\text{recon}}$ is a reconstruction loss that measures the similarity between the two videos. o_i is the initial frame of v_i .

4.2 Phase 2: Learning Spatio-Temporal Constraints

After SFT, the generative foundation model can imagine embodied sequential observations resulting from aerial motion intentions. However, for the unique motion scenarios of 6DoF aerial agents, the predicted outcomes remain unstable. Specifically, spatial inconsistencies arise, such as objects with unrealistic shapes (e.g.,

cars appearing circular) or implausible spatial relationships (e.g., roads floating in the air). Temporally, unnatural deformations occur, such as a pedestrian suddenly splitting into two or buildings continuously twisting. Can the prediction quality of the model be further improved under limited training data? We propose a self-play training process, which involves generating synthetic data pairs incorporating a spatio-temporal discriminator.

a. Motion Intention Generation. Firstly, we randomly sample a video from the original training dataset and then randomly select a frame from the video. This frame is used as the current observation for the aerial agent o . We design a motion prompt p_{LLM} for the LMM to mimic the intentions present in the training dataset and randomly generate a basic intention p_0 . Next, the LMM is tasked with expanding the basic intention, generating M extended intentions $p_1, p_2, \dots, p_M$, ranging from concise to complex, in the following format:

Basic: subject + intention
Extended: subject + intention [intention description] + potential outcomes [future observation description].

The key insight here is that we aim to obtain multiple linguistic expressions $\{p_j\}_{j=0}^M$ for the same intention. By leveraging the multimodal understanding and text generation capabilities of the LMM, we can generate coherent and reasonable textual intentions:

$$\{p_j\}_{j=0}^M = \text{LMM}(o | p_{\text{LLM}}). \quad (3)$$

b. Video Generation. For each intention p_j and the condition observation o , the world model W generates multiple videos $\{v_{j,k}\}_{k=1}^K$ using different random seeds s_k :

$$\hat{v}_{j,k} = W(o, p_j | s_k), \quad j = 0, \dots, M, \quad k = 1, \dots, K. \quad (4)$$

This results in a set of candidate videos $\{\hat{v}_{j,k}\}$ for the similar motion intention of the aerial agent.

c. Rejection Sampling. We aim to design a discriminator to identify which video, among multiple inputs with similar motion intentions, better satisfies spatio-temporal constraints. We propose the following four features, which reflect the quality of predicted videos from different perspectives:

- **Intention Alignment x' :** This feature evaluates whether the generated video aligns with the intended motion by analyzing differences in implicit trajectories across videos $\{\hat{v}_{j,k}\}$. First, the 3D environment and trajectory coordinates are reconstructed from each video. An anomaly detection algorithm is then applied to filter out abnormal trajectories. The underlying assumption is that most generated videos adhere to the motion intention, while a small number of divergent motions can be identified. Specifically, we use VGGT [60] to extract trajectories and isolation forest [41] to detect anomalous trajectories.
- **Temporal Continuity x'' :** The states of objects should change continuously over time, without abrupt jumps or discontinuities. In this case, the short-term observation between consecutive frames is approximately linear. Thus, we extract the even-numbered frames from the video and synthesize them by averaging the adjacent odd-numbered frames. The smoothness is then assessed by calculating the Mean Absolute Error between the real and synthesized even-numbered frames [40].
- **Dynamic Degree x''' :** Video foundation models tend to generate results with minimal or no movement [32]. In aerial motion scenarios, we expect the world model to produce actions with relatively larger motion amplitudes. RAFT [54] is employed to evaluate the dynamics of the generated embodied observations.
- **Spatial Rationality x'''' :** The generated observations often exhibit chaotic and unrealistic spatial structures, such as distorted buildings or large patches of snow, which are particularly prominent in the final frames. We adapt two pre-trained models, LAION [37] and MUSIQ [34], to assess the quality of the final frame, thereby inferring the spatial rationality of the video.

The above process can be summarized as a feature extractor G :

$$\{x'_{j,k}, x''_{j,k}, x'''_{j,k}, x''''_{j,k}\} = G(\hat{v}_{j,k}) \quad (5)$$

We then manually annotated a dataset, selecting videos that best satisfy spatiotemporal constraints, denoted as $\mathcal{D}_{\text{discriminator}}$. We further train a machine learning model F using the four aforementioned features for fitting.

By employing F to output scores for each video, we can obtain a video that aligns with the basic intention p_0 and satisfies spatio-temporal constraints:

$$v^* = \arg \max_{v_{j,k}} F(G(v_{j,k})), \quad (6)$$

d. Synthetic Data Collection. The selected video v^* and its corresponding basic intention prompt p_0 form a synthetic data pair (v^*, p_0) . This pair is added to the synthetic dataset $\mathcal{D}_{\text{synthetic}}$.

$$\mathcal{D}_{\text{synthetic}} \leftarrow \mathcal{D}_{\text{synthetic}} \cup \{(v^*, p_0)\}. \quad (7)$$

e. Supervised Fine-Tuning. When the size of the synthetic dataset $\mathcal{D}_{\text{synthetic}}$ reaches the predefined threshold, it is used to further train the world model W . The training objective is similar to the fine-tuning phase, where the reconstruction loss is minimized:

$$\mathcal{L}_{\text{self-play}} = \frac{1}{|\mathcal{D}_{\text{synthetic}}|} \sum_{(v,p) \in \mathcal{D}_{\text{synthetic}}} \mathcal{L}_{\text{recon}}(W(o, p), v). \quad (8)$$

In this process, the critics of the discriminator are utilized to extract high-quality predictions from the world model. These predictions are then enhanced during SFT, while generations that violate spatio-temporal constraints are suppressed, ultimately enabling the prediction of future observations for 6DoF aerial agents.

5 Experiments

5.1 Experimental Setup

Implementation Details. The proposed dataset is randomly divided into training and testing sets with a ratio of 9:1. We build AirScape based on the video generation foundation model CogVideoX-i2v-5B [66], with main training parameters set as follows: a video resolution of 49×480×720 (frames×height×width), a batch size of 2, gradient accumulation steps of 8, and a total of 10 training epochs. The model was trained on 8 NVIDIA A800-SXM4-40GB GPUs. Additionally, we employed the VLM model Gemini-2.0-Flash [2] in the Phase 2 intention generation, which was selected for its superior video understanding capabilities and efficient response speed. The size of $\mathcal{D}_{\text{discriminator}}$ is 500 video groups, each containing 8–16 videos. The machine learning model F is implemented via Random Forest [6].

Metrics. We evaluate the quality of the world model's predictive embodied observations from two perspectives: (1) the spatio-temporal distribution differences between the generated videos and the ground truth, and (2) the semantic alignment between the generated videos and the input intention.

- **Automatic Evaluation:** FID [27] is used to measure the frame-wise distribution differences between the generated videos and the ground truth videos. For FID evaluation, we crop and resize the predicted frames to match the resolution of the ground truth. FVD [56] evaluates the distribution differences in the temporal dimension. For FVD evaluation, all generated videos and ground truth videos are uniformly downsampled to the same number of frames.
- **Human Evaluation:** The counterfactual reasoning capability of the world model requires assessing whether the generated future observations align with the potential outcomes caused by the motion intention. This evaluation involves judging the semantic consistency between the generated videos and the input intention. However, even state-of-the-art video understanding models struggle to accurately capture the semantic relationship between embodied observations and actions [72]. Following the evaluation ideas in [4, 21, 61], we opt for human evaluation for a more reliable analysis. Specifically, participants are presented with the intention input and the corresponding generated video, and are asked to judge whether they are semantically aligned or not (binary choice). The average intention alignment rate (IAR) is then calculated across the entire test set. For each method, approximately 1.1k generated videos are evaluated, which are randomly and evenly distributed among 9 participants for rating.

Table 1: Evaluation results of predictive future sequence observations for 6DoF aerial agents in three-dimensional space.

Model	Translation			Rotation			Compound			Average		
	FID ↓	FVD ↓	IAR/% ↑	FID ↓	FVD ↓	IAR/% ↑	FID ↓	FVD ↓	IAR/% ↑	FID ↓	FVD ↓	IAR/% ↑
Video Generation Foundation Model												
LTX-Video-2B [26]	153.42	3576.48	37.10	164.52	1097.05	23.53	153.45	1002.81	19.05	154.72	2600.90	26.56
CogVideoX-I2V-5B [66]	126.24	2656.45	30.61	153.96	733.68	27.27	121.34	895.32	14.55	127.89	1947.60	24.14
HunyuanVideo-I2V [36]	173.03	1423.45	22.41	216.97	614.52	8.33	189.77	343.73	35.29	182.60	1043.38	22.01
Wan2.1-I2V-14B [57]	150.46	2622.07	32.35	183.85	1003.68	28.57	165.89	1134.29	24.00	158.47	2036.52	28.31
World Foundation Model												
Cosmos-Predict2-2B-Video2World [45]	236.71	2496.94	22.81	255.74	942.57	33.33	234.82	949.99	29.03	238.43	1903.08	28.39
Cosmos-Predict1-7B-Video2World [1]	142.52	2840.73	36.21	159.60	1171.47	26.92	142.43	1263.72	32.31	144.48	2225.45	31.81
Aerial World Model												
AirScape	104.07	824.75	84.44	142.67	623.53	81.82	114.19	468.49	87.27	111.16	701.90	84.51

**Figure 5: Case analysis of our AirScape and baseline methods, highlighting three common generation issues: limited motion amplitude, shape distortion of spatial objects, and temporal discontinuity.**

Baselines. Due to the absence of world models designed for aerial agents, direct comparisons are not feasible. The most relevant baselines fall into two categories: video generation foundation models and world foundation models. The former includes four popular models: LTX-Video-2B [26], CogVideoX-I2V-5B [66], HunyuanVideo-I2V [36], and Wan2.1-I2V-14B [57]. The latter comprises two different versions of the Cosmos-Predict world models. These models span a parameter range of 2B to 14B.

5.2 Quantitative Results

We present the experimental results of model performance in Table 1. We consider three groups of comparisons, based on which we have the following conclusions.

- **Our proposed AirScape achieves the overall best performance.** Compared with the best-performing baseline models

across the three metrics, AirScape outperforms them with average improvements of 15.47%, 32.73%, and 52.7% on FID, FVD, and IAR metrics, respectively. It is worth mentioning in the Rotation group which requires high ability of physical law following, the AirScape’s performance gain compared with the best baselines are significant.

- **Outcome prediction of 3D aerial motion is challenging.** We can find that, although HunyuanVideo-I2V achieves the best performance on one metric, FVD, in the Compound group, it has poor performance on FID metric in the same group. This phenomenon also exists for other baselines, indicating that the generation focus and optimization direction of existing baseline models are misaligned. Our AirScape achieves overall stable performance, which further verifies the effectiveness of our design.

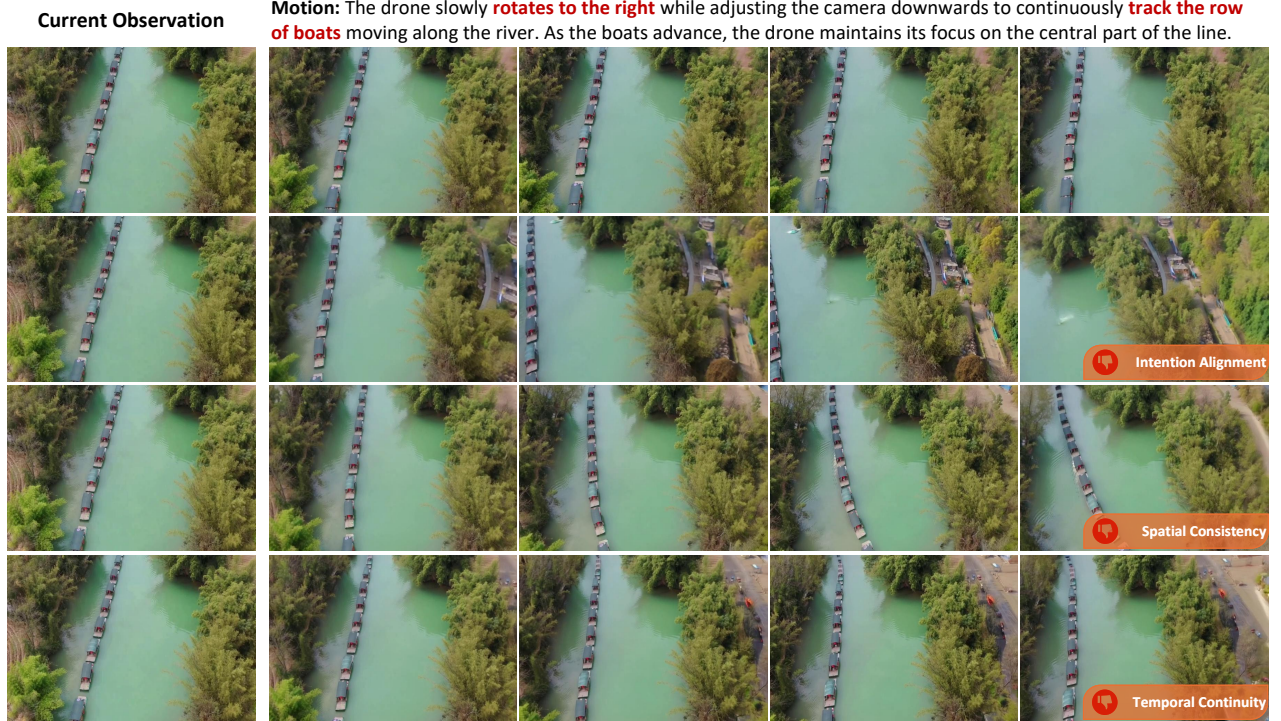


Figure 6: In Phase 2 training, different videos are generated under the same basic intention through rollouts. The spatio-temporal discriminator selects the outcome that best aligns with the intention while satisfying physical spatio-temporal constraints.

- **Model size is not necessarily perfectly correlated with performance.** While larger models generally produce better results, some smaller models can also achieve competitive outcomes. This suggests that there is still room for improving state-of-the-art methods. With carefully curated datasets and advanced training techniques, model performance can be further enhanced.

5.3 Case Analysis

As shown in Figure 5, we present an example illustrating the results generated by the baseline models and AirScape. The output from CogVideoX-I2V-5B appears nearly static, indicating a lack of understanding of motion intention. The results from HunyuanVideo-I2V exhibit distortion in the lower-left region of the final frames, which violates spatial physical consistency. For Cosmos-Predict1-7B-Video2World, the white building on the right undergoes abrupt changes in the temporal sequence, failing to maintain temporal continuity. In contrast, the proposed AirScape effectively predicts sequence observations under motion intention while adhering to spatio-temporal constraints.

5.4 How the Self-Play Works?

The rolled-out videos exhibit noticeable differences, as shown in Figure 6. By increasing the number of generations for each input, we aim to ensure that at least one video sample with good prediction quality is generated. The spatio-temporal discriminator evaluates whether videos satisfy the motion intention alignment and spatio-temporal constraints of embodied observations. In the second row of videos in Figure 6, the row of boats disappears from the frame in later stages, violating the intended goal. In the third row, the row of boats undergoes unnatural distortion, breaking spatial consistency.

Table 2: Ablation study of two-phase training schedule.

Training	FID ↓		FVD ↓	
	Avg.	Std.	Avg.	Std.
After Phase 1	110.98	722.95	59.56	1097.18
After Phase 2	111.16	701.90 $\downarrow 2.9\%$	57.78	1044.47 $\downarrow 4.8\%$

In the fourth row, a hill suddenly appears in the right field of view, disrupting temporal continuity. The first row of videos has the highest rating among the examples. The above process yields a high-quality synthesized dataset.

After further training on the synthesized dataset, the stability of the model’s prediction quality improved. As shown in Table 2, the standard deviations of FID and FVD decreased by 2.9% and 4.8%, respectively, reducing violations of spatio-temporal constraints in the predictions.

6 Conclusion and Future Work

This paper introduces the first aerial world model capable of imagining future embodied observational sequences based on motion intentions. We present a dataset comprising 11k video-motion pairs and a two-phase training schedule for foundation models. Experimental results reveal that aerial spatial imagination poses significant challenges to existing models, while our proposed AirScape achieves substantial improvements across all metrics. In the future, we aim to enhance 1) real-time performance, 2) lightweight design, and 3) applicability for assisting decision-making in real-world aerial agent operations.

Acknowledgments

This paper was supported by the Natural Science Foundation of China under Grant 62371269, Shenzhen Science and Technology Plan Project KJZD20240903102700001, Meituan Academy of Robotics Shenzhen, Talent Program of Guangdong Province (2021QN02Z1 07), National Science and Technology Major Project (2024ZD01NL00 103) and the Major Key Project of PCL (PCL2025A03).

References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. 2025. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575* (2025).
- [2] Google AI. 2023. Gemini-2.0-Flash. <https://ai.google.dev/>. Accessed: 2025-05-30.
- [3] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. 2024. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
- [5] 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).
- [6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [7] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.
- [8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.
- [9] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2024. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [10] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*. PMLR, 1691–1703.
- [11] Xuecheng Chen, Haoyang Wang, Yuhuan Cheng, Hao hao Fu, Yuxuan Liu, Fan Dang, Yunhao Liu, Jinjiang Cui, and Xinlei Chen. 2024. Ddl: Empowering delivery drones with large-scale urban sensing capability. *IEEE Journal of Selected Topics in Signal Processing* (2024).
- [12] Xuecheng Chen, Zijian Xiao, Yuhuan Cheng, Chen-Chun Hsia, Haoyang Wang, Jingao Xu, Susu Xu, Fan Dang, Xiao-Ping Zhang, Yunhao Liu, et al. 2024. Soscheduler: Toward proactive and adaptive wildfire suppression via multi-uav collaborative scheduling. *IEEE Internet of Things Journal* 11, 14 (2024), 24858–24871.
- [13] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476* (2024).
- [14] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 75–1.
- [15] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.
- [16] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. 2019. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215* (2019).
- [17] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. 2024. Understanding world or predicting future? a comprehensive survey of world models. *Comput. Surveys* (2024).
- [18] Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, et al. 2024. EmbodiedCity: A Benchmark Platform for Embodied Agent in Real-world City Environment. *arXiv preprint arXiv:2410.09604* (2024).
- [19] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. 2024. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475* (2024).
- [20] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. 2023. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601* (2023).
- [21] Shenyan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398* (2024).
- [22] Xinyang Gu, Yen-Jen Wang, Xiang Zhu, Chengming Shi, Yanjiang Guo, Yichen Liu, and Jianyu Chen. 2024. Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning. *arXiv preprint arXiv:2408.14472* (2024).
- [23] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Guohui Zhang, and Chengzhong Xu. 2024. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles* (2024).
- [24] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied intelligence via learning and evolution. *Nature communications* 12, 1 (2021), 5721.
- [25] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122* (2018).
- [26] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. 2024. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103* (2024).
- [27] 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).
- [28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851. https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
- [29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in neural information processing systems* 35 (2022), 8633–8646.
- [30] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022).
- [31] Yaosi Hu, Chong Luo, and Zhenzhong Chen. 2022. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18219–18228.
- [32] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- [33] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. 2024. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6689–6700.
- [34] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5148–5157.
- [35] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vignesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125* (2023).
- [36] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuo Zhuo 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).
- [37] LAION-AI. 2022. aesthetic-predictor. <https://github.com/LAION-AI/aesthetic-predictor>.
- [38] Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* 62, 1 (2022), 1–62.
- [39] Chenhao Li, Andreas Krause, and Marco Hutter. 2025. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100* (2025).
- [40] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. 2023. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9801–9810.
- [41] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 eighth IEEE international conference on data mining*. IEEE, 413–422.
- [42] Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C Pérez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, et al. 2024. Mardini: Masked autoregressive diffusion for video generation at scale. *arXiv preprint arXiv:2410.20280* (2024).
- [43] Yuxuan Liu, Haoyang Wang, Fanhang Man, Jingao Xu, Fan Dang, Yunhao Liu, Xiao-Ping Zhang, and Xinlei Chen. 2024. Mobair: Unleashing sensor mobility for city-scale and fine-grained air-quality monitoring with Airbert. In *Proceedings*

- of the 22nd Annual International Conference on Mobile Systems, Applications and Services. 223–236.
- [44] Zhishuo Liu, Xingquan Zuo, Mengchu Zhou, Bin Jia, and Chongyang Xin. 2025. Meal Delivery Routing Problem with a Hybrid Fleet of Riders and Autonomous Vehicles under Dynamic Environment. *IEEE Transactions on Automation Science and Engineering* (2025).
 - [45] NVIDIA. 2025. COSMOS Predict2. <https://github.com/nvidia-cosmos/cosmos-predict2>. Accessed: 2025-08-26.
 - [46] OpenAI. 2023. Sora: High-Fidelity Video Generation. <https://openai.com/sora/>. Accessed: 2025-05-30.
 - [47] Robert Panowicz and Wojciech Stecz. 2024. Robust Optimization Models for Planning Drone Swarm Missions. *Drones* 8, 10 (2024), 572.
 - [48] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
 - [49] Sandra Rafael, Luís P Correia, Diogo Lopes, Jorge Bandeira, Margarida C Coelho, Mário Andrade, Carlos Borrego, and Ana I Miranda. 2020. Autonomous vehicles opportunities for cities air quality. *Science of the Total Environment* 712 (2020), 136546.
 - [50] 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).
 - [51] Ryo Sakagami, Florian S Lay, Andreas Dömel, Martin J Schuster, Alin Albu-Schäffer, and Freek Stulp. 2023. Robotic world models—conceptualization, review, and engineering best practices. *Frontiers in Robotics and AI* 10 (2023), 1253049.
 - [52] David C Schedl, Indrajit Kurmi, and Oliver Bimber. 2021. An autonomous drone for search and rescue in forests using airborne optical sectioning. *Science Robotics* 6, 55 (2021), eabg1188.
 - [53] Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. 2024. Hierarchical patch diffusion models for high-resolution video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7569–7579.
 - [54] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*. Springer, 402–419.
 - [55] Marko Tot, Shu Ishida, Abdelhak Lemkhenter, David Bignell, Pallavi Choudhury, Chris Lovett, Luis França, Matheus Ribeiro Furtado de Mendonça, Tarun Gupta, Darren Gehring, et al. 2025. Adapting a World Model for Trajectory Following in a 3D Game. *arXiv preprint arXiv:2504.12299* (2025).
 - [56] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
 - [57] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).
 - [58] Haoyang Wang, Jingao Xu, Xinyu Luo, Xuecheng Chen, Ting Zhang, Ruiyang Duan, Yunhao Liu, and Xinlei Chen. 2025. Ultra-high-frequency harmony: mmwave radar and event camera orchestrate accurate drone landing. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. 15–29.
 - [59] Haoyang Wang, Jingao Xu, Chenyu Zhao, Zihong Lu, Yuhao Cheng, Xuecheng Chen, Xiao-Ping Zhang, Yunhao Liu, and Xinlei Chen. 2024. Transformloc: Transforming mavs into mobile localization infrastructures in heterogeneous swarms. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 1101–1110.
 - [60] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5294–5306.
 - [61] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. 2023. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. (2023).
 - [62] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. 2024. DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving. In *European Conference on Computer Vision*. Springer, 55–72.
 - [63] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. 2024. WorldDreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985* (2024).
 - [64] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. 2023. Daydreamer: World models for physical robot learning. In *Conference on robot learning*. PMLR, 2226–2240.
 - [65] Yanggang Xu, Jirong Zha, Jiyuan Ren, Xintao Jiang, Hongfei Zhang, and Xinlei Chen. 2024. Scalable multi-agent reinforcement learning for effective uav scheduling in multi-hop emergency networks. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 2028–2033.
 - [66] 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).
 - [67] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. 2022. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8896–8905.
 - [68] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.
 - [69] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786* (2025).
 - [70] Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yanan Yang, Xiang Wan, Shiming Ge, and Dacheng Tao. 2023. WebUAV-3M: A Benchmark for Unveiling the Power of Million-Scale Deep UAV Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2023), 9186–9205. doi:10.1109/TPAMI.2022.3232854
 - [71] 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. *arXiv preprint arXiv:2505.05622* (2025).
 - [72] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyu Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025. UrbanVideoBench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces. 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, 32400–32423. doi:10.18653/v1/2025.acl-long.1558
 - [73] Baining Zhao, Ziyu Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. 2025. Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning in Foundation Models via Reinforcement Learning. *arXiv:2504.12680 [cs.AI]* <https://arxiv.org/abs/2504.12680>
 - [74] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. 2024. Identifying and solving conditional image leakage in image-to-video diffusion model. *arXiv preprint arXiv:2406.15735* (2024).
 - [75] Nan Zhou, Yuxuan Liu, Haoyang Wang, Fanhang Man, Jingao Xu, Fan Dang, Chaopeng Hong, Yunhao Liu, Xiao-Ping Zhang, Yali Song, et al. 2025. CatUA: Catalyzing Urban Air Quality Intelligence through Mobile Crowd-sensing. *IEEE Transactions on Mobile Computing* (2025).
 - [76] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. 2024. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377* (2024).