AETHER: Geometric-Aware Unified World Modeling

Anonymous ICCV RIWM submission

Paper ID

Figure 1. **An overview of AETHER, trained entirely on synthetic data.** The figure highlights its three key capabilities: 4D reconstruction, action-conditioned 4D prediction, and visual planning, all demonstrated on unseen real-world data. The 4D reconstruction examples are derived from MovieGen [48] and Veo 2 [62] generated videos, while the action-conditioned prediction uses an observation image from a university classroom. The visual planning example utilizes observation and goal images from an office building. Better viewed when zoomed in. Additional visualizations can be found on our website.

Abstract

The integration of geometric reconstruction and generative modeling remains a critical challenge in developing AI systems capable of human-like spatial reasoning. This paper proposes AETHER, *a unified framework that enables geometry-aware reasoning in world models by jointly optimizing three core capabilities: (1) 4D dynamic reconstruction, (2) action-conditioned video prediction, and (3) goal-conditioned visual planning. Through task-interleaved feature learning,* AETHER *achieves synergistic knowledge sharing across reconstruction, prediction, and planning objectives. Building upon video generation models, our framework demonstrates zero-shot synthetic-to-real generalization despite never observing real-world data during training. Furthermore, our approach achieves zero-shot generalization in both action following and reconstruction tasks, thanks to its intrinsic geometric modeling. Notably, even without real-world data, its reconstruction performance is comparable with or even better than that of domain-specific models. Additionally,* AETHER *employs camera trajectories as geometry-informed action spaces, enabling effective action-conditioned prediction and visual planning. We hope our work inspires the community to explore new frontiers in physically-reasonable world modeling and its applications.*

1. Introduction

"Prediction is not just one of the things your brain does. It is the primary function of the neocortex."

— Jeff Hawkins, *On Intelligence* (2004)

The development of visual intelligence systems capable of comprehending and forecasting the physical world remains a cornerstone of AI research. World models have emerged as a foundational paradigm for building autonomous systems that not only perceive but also anticipate environmental dynamics to make reasonable actions. At their core, three capabilities stand out: First, perception equips the system with the ability to capture the intricate four-dimensional (4D) changes—integrating spatial and temporal information—that are essential for understanding the physical world [37, 63, 65, 66, 82, 86]. This continuous sensing of dynamic cues enables a geometric representation of the environment. Second, prediction leverages this perceptual information to forecast how the environment will evolve under specific actions, thereby providing a foresight into future states [3, 24, 28, 32, 35, 60, 77]. Finally, planning uses these predictive insights to determine the optimal sequence of actions required to achieve a given goal. Together, these three aspects empower world models to not only represent the current state of the environment but also to anticipate and navigate its future dynamics effectively.

Motivated by these principles, we introduce AETHER, a unified framework that, for the first time, bridges reconstruction, prediction, and planning, as shown in Fig. 1. AETHER leverages pre-trained video generation models [28, 77] and is further refined via post-training with synthetic 4D data. Although multiple action modalities exist, ranging from keyboard inputs [2, 11, 15, 46, 79] to human or robotic motions [16, 84, 89, 90] and point flows [22, 69], we choose camera pose trajectories as our global action representation. This choice is particularly effective for ego-view tasks: in navigation, camera trajectories directly correspond to the navigation paths, while in robotic manipulation, the movement of an in-hand camera captures the 6D motion of the end effector. To address the scarcity of 4D data, we utilize RGB-D synthetic video data and propose a robust camera pose annotation pipeline to reconstruct full 4D dynamics.

Through a simple training strategy that randomly combines input and output modalities, our method transforms the base video generation model into a unified, multi-task world model with three key capabilities: (1) Depth and camera pose estimation from full video sequences; (2) Video prediction conditioned on an initial observation—with the option to incorporate a camera trajectory action; and (3) Goal-conditioned visual planning based on observation–goal image pairs. We transform depth videos into scale-invariant normalized disparity representations to meet the tokenization requirements of video VAEs. Simultaneously, we encode camera trajectories as scale-invariant raymap sequence representations, structured to align with the spatiotemporal framework of diffusion transformers (DiTs). By dynamically integrating cross-task and cross-modal conditioning signals during training, our framework enables synergistic knowledge transfer across heterogeneous inputs, facilitating joint optimization for multi-task generative modeling.

In summary, this work introduces AETHER, a unified world model that integrates reconstruction, prediction, and planning through multi-task learning on synthetic 4D data. We propose a robust automatic data annotation pipeline to obtain accurate 4D geometry knowledge. By combining geometric reasoning with generative priors, our framework achieves robust zero-shot transfer to real-world tasks, demonstrating accuracy comparable to SOTA reconstruction models while enabling actionable planning capabilities. The results underscore the value of synergistic 4D modeling for advancing spatial intelligence in AI systems. We hope that AETHER will serve as an effective starter framework for the community to explore post-training world models with scalable synthetic data.

2. 4D Synthetic Data Annotation Pipeline

For the synthetic data source, we follow DA-V [74] and TheMatrix [17] to collect large-scale synthetic data with high-quality video depth data. With high-resolution RGB videos

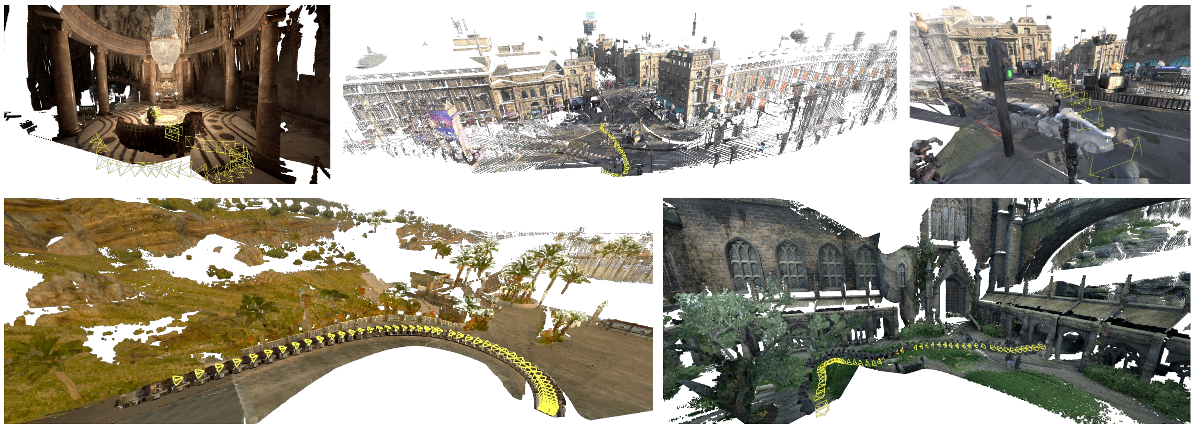

Figure 2. Some visualization results of data annotated through our pipeline. Better viewed when zoomed in.

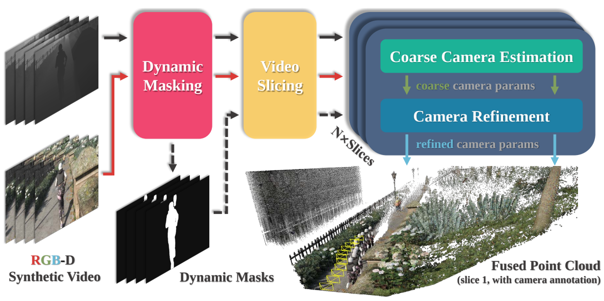

Figure 3. Our robust automatic camera annotation pipeline.

and corresponding per-frame depth maps collected, we built a **robust** and fully **automatic** camera annotation pipeline for both camera extrinsics and intrinsics. As illustrated in Fig. 3, the pipeline has four stages: (1) object-level dynamic masking, (2) reconstruction-friendly video slicing, (3) coarse camera localization and calibration, and (4) tracking-based camera refinement with bundle adjustment. We present several visualizations of our annotated data in Fig. 2, ranging from indoor to outdoor scenes, and from static to dynamic scenarios, demonstrating the robustness and accuracy of our annotation method.

Dynamic Masking. Distinguishing between dynamic and static regions is crucial for accurate camera parameters estimation. Here, we utilize semantic categories that are potentially dynamic (e.g., cars, people) to segment dynamic objects. Although this may occasionally misclassify static objects, such as stationary parked cars, as dynamic, we find it more robust than flow-based segmentation methods. Specifically, we use Grounded SAM 2 [50] to ensure the temporal consistency of dynamic masks over long sequences.

Video Slicing. Video slicing plays a critical role in 3D reconstruction by serving two key purposes: First, it eliminates unsuitable video segments (such as scene cuts or motion-blurred frames) that could compromise reconstruction quality. Second, it segments long videos into shorter, temporally coherent clips to enhance processing efficiency. The specific criteria for frame removal are as follows: (1) *Insufficient Feature Points*: We employ the SIFT [39] feature descriptor to extract keypoints from each frame. Frames exhibiting insufficient SIFT keypoints are discarded to ensure robust correspondence estimation. Additionally, frames containing regions with insufficient texture due to low illumination are excluded, as such areas typically exhibit poor feature discriminability and pose challenges for reliable matching. (2) *Large Areas of Dynamic Regions*: Frames where dynamic regions (obtained from dynamic annotation) dominate over static regions can introduce ambiguity in camera pose estimation. Such frames are filtered out to ensure robust results. (3) *Large Motion or Inaccurate Correspondence*: Using an off-the-shelf optical flow estimator, RAFT [61], we estimate the magnitude of motion. If these magnitudes exceed a predefined threshold, we truncate the sequence at the current frame, retaining all preceding frames as a valid segment. Similarly, if the ratio of forward-to-backward optical flow errors surpasses a threshold value, we truncate the current frames to ensure temporal coherence.

Coarse Camera Estimation. For each video slice, we first use DroidCalib [25] to perform a coarse estimation of the camera parameters, leveraging the depth information from static regions. However, due to the lower input resolution of the DroidCalib model and the limited accuracy of its correspondence estimation, a refinement process is necessary to obtain precise camera parameters.

Camera Refinement. We begin camera refinement by employing the state-of-the-art tracker, CoTracker3 [33], to capture accurate long-term correspondences across the entire slice. SIFT [39] and SuperPoint [12] feature points are extracted from static regions, and then tracked to form correspondences. Subsequently, bundle adjustment is performed on all frames to minimize the accumulated reprojection error of all correspondences. With access to high-quality dense depth, we apply forward-backward reprojection to estimate and minimize errors in 3D space [8], which improves per-frame camera accuracy while preserving inter-frame geometric consistency. Specifically, we solve the nonlinear optimization problem by Ceres Solver [1], and the Cauchy loss function is applied to measure correspondence residuals, which accounts for the problem's sparsity.

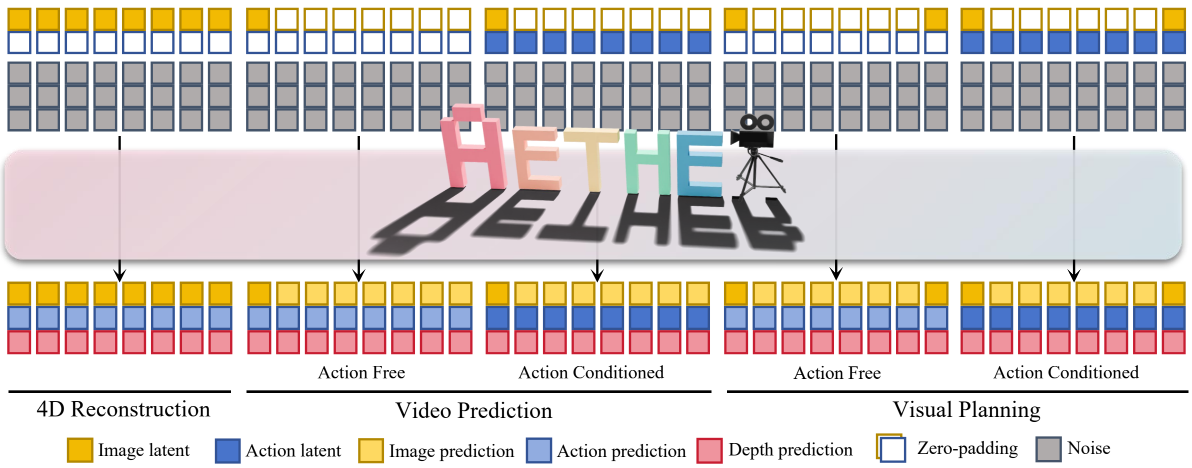

Figure 4. The overall pipeline of AETHER. With different condition combinations, AETHER can serve different tasks.

3. AETHER Multi-Task World Model

In this section, we introduce how we post-train a base video diffusion model into a unified multi-task world model AETHER. We use CogVideoX-5b-I2V [77] as our base model. We first give an overview of our framework in Sec. 3.1, then we detail on the input process of depth videos and camera pose trajectories in Sec. 3.2 and Sec. 3.3. Finally, we show how we do model training in Sec. 3.4.

3.1. Method Overview

Mainstream video diffusion models [27, 40] typically involve two processes: a forward (noising) process and a reverse (denoising) process. The forward process incrementally adds Gaussian noise, denoted as $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, to a clean latent sample $\mathbf{z}_0 \in \mathbb{R}^{k \times c \times h \times w}$, where k, c, h, w represent the dimensions of the video latents. Through this process, the clean $\mathbf{z}_0$ is gradually transformed into a noisy latent $\mathbf{z}_t$. In the reverse process, a learned denoising model ϵ_θ progressively removes the noise from $\mathbf{z}_t$ to reconstruct the original latent representation. The denoising model ϵ_θ is conditioned on auxiliary inputs $\mathbf{c}$ and the diffusion timestep t.

In our method, the target latent $\mathbf{z}_0$ comprises three modalities: color video latents $\mathbf{z}_{\mathrm{c}0}$, depth video latents $\mathbf{z}_{\mathrm{d}0}$, and action latents $\mathbf{z}_{\mathrm{a}0}$. The model additionally takes two types of conditions as input: color video conditions $\mathbf{c}_{\mathrm{c}}$ and action conditions $\mathbf{c}_{\mathrm{a}}$. For the action modality, we choose *camera pose trajectory* as a global action, facilitated by our automated camera pose annotation pipeline described earlier. All latents and conditions are channel-wise concatenated. The training objective of AETHER can be expressed as:

$$\mathcal{L}_\theta = \mathbb{E}_{\substack{\epsilon \sim \mathcal{N}(0,\mathbf{I}) \\ t \sim \mathcal{U}(1,\mathcal{T}) \\ \mathbf{z}_0 = \mathbf{z}_{\mathrm{c}0} \otimes \mathbf{z}_{\mathrm{d}0} \otimes \mathbf{z}_{\mathrm{a}0} \\ \mathbf{c} = \mathbf{c}_{\mathrm{c}} \otimes \mathbf{c}_{\mathrm{a}}}} \left\| \epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) \right\|^2, \tag{1}$$

where $\otimes$ denotes the channel-wise concatenation operation, $\mathcal{U}(\cdot)$ represents a uniform distribution, and $\mathcal{T}$ denotes the denoising steps.

The multi-task objective of AETHER is determined by the specific conditions $\mathbf{c}$ for different tasks. (1) *Reconstruction*: $\mathbf{c}_{\mathrm{c}}$ represents the input video latents. (2) *Video prediction*: $\mathbf{c}_{\mathrm{c}}$ takes the latent of observation image as the first frame, while other latents are zero-masked. (3) *Goal-conditioned visual planning*: The first and last latents of $\mathbf{c}_{\mathrm{c}}$ correspond to the observation and goal images, respectively, with all intermediate latents zero-padded. For the action condition $\mathbf{c}_{\mathrm{a}}$, it is either entirely zero-masked or contains the full target camera pose trajectory in action-free or action-conditioned control cases. Illustrations are show in Fig. 4.

3.2. Depth Videos Process

Given a depth video $\mathbf{x}_d$, we first clip the depth values to a predefined range $[d_{\min}, d_{\max}]$. Next, we apply a square root transformation and subsequently compute the reciprocal to convert the depth values into disparity, as described in [57]. Each disparity video clip is then normalized in a scale-invariant manner. Subsequently, the normalized disparity values are linearly mapped from $[0, 1]$ to $[-1, 1]$. To meet the input requirements of the VAE, the single-channel disparity map is replicated across three channels, as done in prior works [34, 74]. The final depth latent is computed as:

$$\mathbf{x}_{\mathrm{disp}} = \frac{1}{\sqrt{\mathrm{clip}(\mathbf{x}_d, d_{\min}, d_{\max})}}, \tag{2}$$

$$\hat{\mathbf{x}}_{\mathrm{disp}} = \frac{\mathbf{x}_{\mathrm{disp}}}{\max(\mathbf{x}_{\mathrm{disp}})} \times 2 - 1, \tag{3}$$

$$\mathbf{z}_d = \mathcal{E}(\hat{\mathbf{x}}_{\mathrm{disp}} \otimes \mathbf{1}_3), \tag{4}$$

where $\mathcal{E}$ denotes the 3D VAE, and $\otimes \mathbf{1}_3$ represents the channel-wise replication of 3 times. The above operations are designed to be compatible with the pretrained 3D VAE model, ensuring minimal reconstruction error.

3.3. Camera Trajectories Process

We transform camera parameters into raymap videos [7] so that video diffusion can process them compatibly. Specifically, given the intrinsic matrix $\mathbf{K} \in \mathbb{R}^{T\times3\times3}$ and the extrinsic matrix $\mathbf{E} \in \mathbb{R}^{T\times4\times4}$, the transformation process can be described as follows.

Translation Scaling and Normalization. The translation component of the camera pose (inverse of extrinsic matrix), $\mathbf{t} \in \mathbb{R}^3$, is first scaled by a constant factor s_{ray} and normalized using the maximum disparity value $d_{\max}$. To suppress large values, we then pass it through a signed $\log(1+\cdot)$ transformation:

$$\mathbf{t}' = \frac{\mathbf{t}}{\max\left(x_{\text{disp}}\right)} \cdot s_{\text{ray}}, \tag{5}$$

$$\mathbf{t}_{\log} = \text{sign}(\mathbf{t}') \cdot \log(1 + |\mathbf{t}'|), \tag{6}$$

where s_{ray} is a predefined scaling factor.

Raymap Construction. Using the intrinsic matrix $\mathbf{K}$, we compute the camera ray directions $\mathbf{r}_{\text{d}}$ in homogeneous coordinates for each pixel. Note that we do *not* unit normalize it but let it have a unit value along z axis. The ray origins $\mathbf{r}_{\text{o}}$ are set to the translation $\mathbf{t}_{\log}$, replicated across the spatial dimensions. The raymap in the world coordinate system is obtained by transforming the ray directions $\mathbf{r}_{\text{d}}$ using the extrinsic matrix $\mathbf{E}$. The final raymap $\mathbf{r}$ consists of 6 channels: 3 for the ray directions $\mathbf{r}_{\text{d}}$ and 3 for the ray origins $\mathbf{r}_{\text{o}}$.

Resolution Downsampling. To align the raymap with the latent feature dimensions from the VAE, we perform adjustments both spatially and temporally. Spatially, the raymap is downsampled by a factor of 8 using bilinear interpolation. Temporally, every consecutive group of 4 frames is concatenated along the channel dimension. The resulting rearranged tensor is denoted as $\mathbf{z}_a$

Converting raymap back to camera matrix. Given generated raymap sequences rearranged by the time axis $\hat{\mathbf{r}} \in \mathbb{R}^{T\times6\times h\times w} = [\hat{\mathbf{r}_{\text{d}}}, \hat{\mathbf{r}_{\text{o}}}]$, we first recover the ray origins by:

$$\hat{\mathbf{r}_{\text{o}}}' = \frac{1}{s_{\text{ray}}} \cdot \text{sign}(\hat{\mathbf{r}_{\text{o}}}) \cdot (\exp(|\hat{\mathbf{r}_{\text{o}}}|) - 1), \tag{7}$$

Then, we can recover both the intrinsics and extrinsics through Alg. 1 in the supplementary material.

3.4. Model Training

We initialize AETHER with pre-trained CogVideo-5b-I2V [77] weights, excluding the additional input and output projection layer channels for depth and raymap action trajectories, which are initialized to zero. Since text prompt conditions are not used, an empty text embedding is provided during both training and inference.

As the dataset we use contains video clips with variable lengths and frames per second (FPS), we randomly select $T \in \{17, 25, 33, 41\}$ frames, and the FPS is randomly sampled from $\{8, 10, 12, 15, 24\}$. The RoPE [59] coefficients are linearly interpolated to align with them.

During training, conditional inputs are randomly masked to generalize across tasks. For $\mathbf{c}_c$, masking probabilities are: 30% for both observation and goal images (visual planning tasks), 40% for observation images only (video prediction), 28% for full-color video latents (4D reconstruction), and 2% for masking all of $\mathbf{c}_c$. For $\mathbf{c}_a$, trajectory latents are either kept or fully masked with equal probability (supporting action-free or action-conditioned tasks with raymap conditions). This strategy enables the model to adapt to diverse tasks and input condition settings.

Our training process consists of two stages. In the first stage, we adopt the loss function of a standard latent diffusion model, which minimizes the mean squared error (MSE) in the latent space. In the second stage, we refine the generated outputs by decoding them into the image space. Specifically, we introduce three additional loss terms: a Multi-Scale Structure Similarity (MS-SSIM) loss [67] for color video, a scale- and shift-invariant loss [49] for depth videos, and a scale- and shift-invariant pointmap loss [66] for pointmaps projected from the generated depths and raymaps. Further details on the stage 2 loss functions are provided in the supplementary material. Notably, the second stage takes about $\frac{1}{4}$ of the training steps used in the first stage.

We employ a hybrid training strategy combining Fully Sharded Data Parallel (FSDP) [87] with Zero-2 optimization within compute nodes and Distributed Data Parallel (DDP) across nodes. Since depth videos require online normalization, the VAE encoder is also run online during training and operates under DDP. Our implementation processes a local batch size of 4 per GPU, resulting in an effective batch size of 320 samples across 80 A100-80GB GPUs. Training is conducted over two weeks using the AdamW [38] optimizer with a OneCycle [56] learning rate scheduler.

4. Reconstruction Experiments

In this section, we demonstrate that AETHER can achieve zero-shot reconstruction metrics comparable to or even better than SOTA reconstruction methods. We mainly consider two zero-shot reconstruction tasks: video depth estimation and camera pose estimation. Note that we only denoise for 4 steps for reconstruction tasks.

4.1. Zero-Shot Video Depth Estimation

Implementation Details. Video depth estimation is evaluated based on two key aspects: per-frame depth quality and inter-frame depth consistency. These evaluations are performed by aligning the predicted depth maps with the ground truth using a per-sequence scale. We use absolute relative error (Abs Rel) and $\delta < 1.25$ (percentage of predicted depths within a 1.25-factor of true depth) as metrics. For implementation, we adopt the settings outlined in CUT3R [65]. Our baselines include both reconstruction-based methods—such as DUSt3R [66], MASt3R [37], MonST3R [82], Spann3R [63], and CUT3R [65]—and diffusion-based depth estimators, including ChronoDepth [55], DepthCrafter [29],

Table 1. **Video depth Evaluation.** Methods requiring global alignment are marked “GA”.

Method	Sintel [6]		BONN [44]		KITTI [21]	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
Reconstruction Methods. Alignment: per-sequence scale						
DUSt3R-GA [66]	0.656	45.2	0.155	83.3	0.144	81.3
MASt3R-GA [37]	0.641	43.9	0.252	70.1	0.183	74.5
MonST3R-GA [82]	0.378	**55.8**	**0.067**	**96.3**	0.168	74.4
Spann3R [63]	0.622	42.6	0.144	81.3	0.198	73.7
CUT3R [65]	0.421	47.9	0.078	93.7	0.118	88.1
AETHER (Ours)	**0.324**	50.2	0.273	59.4	**0.056**	**97.8**
Diffusion-Based Methods. Alignment: per-sequence scale&shift						
ChronoDepth [55]	0.429	38.3	0.318	51.8	0.252	54.3
DepthCrafter [29]	0.590	55.5	**0.253**	56.3	0.124	86.5
DA-V [74]	1.252	43.7	0.457	31.1	0.094	93.0
AETHER (Ours)	**0.314**	**60.4**	0.308	**60.2**	**0.054**	**97.7**

Table 2. **Evaluation on Camera Pose Estimation.**

Method	Sintel [6]			TUM-dynamics [58]			ScanNet [10]		
	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓
Optimization-based Methods									
Particle-SfM [86]	0.129	**0.031**	**0.535**	-	-	-	0.136	0.023	0.836
Robust-CVD [36]	0.360	0.154	3.443	0.153	0.026	3.528	0.227	0.064	7.374
CasualSAM [85]	0.141	0.035	0.615	0.071	**0.010**	1.712	0.158	0.034	1.618
DUSt3R-GA [66]	0.417	0.250	5.796	0.083	0.017	3.567	0.081	0.028	0.784
MASt3R-GA [37]	0.185	0.060	1.496	**0.038**	0.012	**0.448**	0.078	0.020	**0.475**
MonST3R-GA [82]	**0.111**	0.044	0.896	0.098	0.019	0.935	**0.077**	**0.018**	0.529
Feed-forward Methods									
DUSt3R [66]	0.290	0.132	7.869	0.140	0.106	3.286	0.246	0.108	8.210
Spann3R [63]	0.329	0.110	4.471	0.056	0.021	0.591	**0.096**	0.023	0.661
CUT3R [65]	0.213	0.066	**0.621**	0.046	0.015	**0.473**	0.099	**0.022**	**0.600**
AETHER (Ours)	**0.189**	**0.054**	0.694	0.092	**0.012**	1.106	0.176	0.028	1.204

and DepthAnyVideo (DA-V) [74]. It is important to note that when comparing with diffusion-based depth estimators, we apply scale and shift alignment to the ground truth, as most of these methods are not inherently scale-invariant. All videos are resized with original aspect ratios kept to make the short side align with our model's input size. For videos that exceed the maximum forward processing spatial or temporal size of our model, we employ a sliding window strategy with a stride size of 8. In regions of overlap between windows, we first estimate a relative scale by calculating the average of element-wise division. This relative scale is then used to adjust the latter window's depth predictions. Finally, a linspace-weighted average is applied to the overlapping areas, following approaches similar to prior methods [29, 80].

Results and Analysis. Table 1 summarizes the video depth estimation results across Sintel [6], BONN [44], and KITTI [21] datasets. For reconstruction-based methods, AETHER outperforms or is comparable with prior approaches. On Sintel, AETHER achieves the lowest Abs Rel (0.324), surpassing MonST3R-GA (0.378), and competitive $\delta < 1.25$ (50.2). On KITTI, AETHER sets a new benchmark with Abs Rel of 0.056 and $\delta < 1.25$ of 97.8, outperforming the previous SOTA CUT3R (Abs Rel: 0.118, $\delta < 1.25$: 88.1). Among diffusion-based methods, AETHER shows consistent superiority. It achieves the best performance on Sintel (Abs Rel: 0.314, $\delta < 1.25$: 60.4) and KITTI (Abs Rel: 0.054, $\delta < 1.25$: 97.7), significantly outperforming ChronoDepth [55], DepthCrafter [29], and DA-V [74]. On BONN, AETHER achieves the highest $\delta < 1.25$ (60.2) with competitive Abs Rel (0.308).

4.2. Zero-Shot Camera Pose Estimation

Implementation Details. Following MonST3R [82] and CUT3R [65], we evaluate camera pose estimation accuracy on the Sintel [6], TUM Dynamics [58], and ScanNet [10] datasets. Notably, both Sintel and TUM Dynamics contain highly dynamic objects, presenting significant challenges for traditional Structure-from-Motion (SfM) and Simultaneous Localization and Mapping (SLAM) systems. We report Absolute Translation Error (ATE), Relative Translation Error (RPE Trans), and Relative Rotation Error (RPE Rot) after Sim(3) alignment with the ground truth, following the methodology in [65]. The implementation settings are consistent with those used in CUT3R [65]. All videos are resized with original aspect ratios kept and then center cropped to align with our model's input size. For long videos exceeding our model's maximum temporal forward processing length, a sliding window strategy with a stride size of 32 is employed. In overlapping regions between windows, camera poses are aligned following prior methods [64]. Translation alignment is performed using linear interpolation, while quaternion rotations are interpolated with spherical linear interpolation. Additionally, we observed that the generated camera trajectories exhibit noise, likely due to the limited number of denoising steps. To mitigate this, we apply a simple Kalman filter [71] to smooth the trajectories.

Results and Analysis. Table 2 shows the evaluation results. Among feed-forward methods, AETHER achieves the best ATE (0.189) and RPE Trans (0.054) on Sintel [6], while remaining competitive in RPE Rot (0.694) compared to CUT3R (0.621). On TUM Dynamics [58], AETHER achieves the best RPE Trans (0.012). For other metrics, AETHER is also comparable with other specialist models.

5. Generation and Planning Experiments

In this section, we first show video prediction, with or without action conditioning, quantitatively or qualitatively, in Sec. 5.1. We then show visual planning abilities in Sec. 5.2. More visualizations are in the supplementary material.

5.1. Video Prediction

Implementation Details. We use CogVideoX-5b-I2V [77] as our baseline. To ensure a fair comparison, we construct a validation dataset comprising two subsets: in-domain and out-domain data. The in-domain subset includes novel, unseen scenes from the same synthetic environments as the training dataset, while the out-domain subset consists of data from entirely new synthetic environments. Both models are provided with the first frame as the observation image. For action-free prediction, since CogVideoX depends heavily on text prompts, we use GPT-4o [31] to generate image descriptions and predictions of future scenes as prompts for CogVideoX. In contrast, AETHER is evaluated using empty text prompts. For action-conditioned prediction, we also labeled camera trajectories in the validation dataset and generated corresponding raymap sequences as action conditions for AETHER. For the baseline, in addition to the prompts used for action-free prediction, we use GPT-4o [31] to generate detailed descriptions of object and camera movements, enabling the baseline to use language as action conditions. We use the default classifier-free guidance value of 6 on text prompts for CogVideoX and a value of 3 on the observation image for AETHER. No classifier-free guidance is applied to action conditions to ensure fairness. Evaluation metrics follow VBench [30], a standard benchmark for video generation, with additional details on prompts and evaluation metrics provided in the supplementary material.

Image-to-Video Prediction. We first evaluate image-to-video prediction without action conditions. The results, presented in Tab. 3, show that AETHER consistently outperforms the baseline on both in-domain and out-domain validation sets. Notably, AETHER demonstrates a larger performance improvement on out-domain data, which can likely be attributed to the baseline model's pre-training data containing domains similar to the in-domain dataset.

Action-Conditioned Video Prediction. To assess the effectiveness of our post-training in improving action control and action-following capabilities, we conduct action-conditioned video prediction experiments. The results, shown in Tab. 4, indicate that AETHER consistently outperforms the baseline in both in-domain and out-domain settings. Notably, CogVideoX tends to generate static scenes with high visual and aesthetic quality, while AETHER accurately follows the action conditions, producing highly dynamic scenes. These results validate the effectiveness of our framework and the advantages of using camera trajectories as action conditions.

5.2. Visual Planning

Implementation Details. We evaluate the action-conditioned navigation capability of AETHER on our validation set. To demonstrate the effectiveness of our multi-task objective, particularly the incorporation of the reconstruction objective, we also post-train an ablation model without the video depth objective, denoted as AETHER-no-depth. Given the observation image, goal image, and camera trajectory, the resulting video should be highly determined. Thus, we report pixel-wise reconstruction metrics, including PSNR, SSIM [68], MS-SSIM [67], and LPIPS [83], for action-conditioned navigation. For the action-free case, which represents a visual path navigation task, we also report the VBench metrics. We do *not* use any classifier-free guidance on both tasks.

Action-Conditioned Navigation. The quantitative results for action-conditioned navigation are presented in Tab. 5. AETHER consistently outperforms the ablation model, demonstrating the significant benefits of incorporating the reconstruction objective into generative models.

Visual Path Planning. In the absence of action conditions, this task evaluates the model's ability to function as a "world model as an agent," requiring it to plan a path from the observation image to the goal image. The results, shown in Tab. 6, indicate that the reconstruction objective substantially improves the model's visual path planning capability. Additionally, qualitative visualizations on completely in-the-wild data are provided in supplementary material.

6. Related Work

World Models. World models have emerged as a critical framework in artificial intelligence, enabling agents to simulate, understand, and predict environmental dynamics. Early work [23] introduced latent representations and recurrent neural networks for decision-making. Recent advancements include Cat3D [20] for 3D scene generation, Cat4D [72] for dynamic 4D environments, and Genie 2 [46], a large-scale model for interactive 3D worlds. Motion Prompting [22] further enables precise video generation control. These ad-

Table 3. **VBench [30] Metrics of Video Prediction without Action Conditions.** Comparison between CogVideoX and AETHER (Ours) on ***in-domain/out-domain/overall*** performance on the validation set. For each group, the better performance is highlighted in **bold**.

	subject consistency	b.g. consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	weighted average
CogVideoX	89.36/84.61/87.77	92.72/91.43/92.29	98.24/96.93/97.81	88.75/95.00/90.83	**54.49/53.58/54.18**	55.38/52.29/54.35	79.01/77.52/78.51
AETHER	**91.50/87.55/90.18**	**94.29/93.62/94.07**	**98.54/98.19/98.42**	**96.25/100.00/97.50**	54.36/52.58/53.77	**55.08/54.88/55.01**	**80.34/79.42/80.04**

Table 4. **VBench [30] Metrics of Action-Conditioned Video Prediction.** Comparison between CogVideoX and AETHER (Ours) on ***in-domain/out-domain/overall*** performance on the validation set. For each metric group, the better performance is highlighted in **bold**.

	subject consistency	b.g. consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	weighted average
CogVideoX	**91.56**/88.23/90.51	92.98/92.29/92.77	98.44/97.81/98.24	83.87/**93.02**/86.76	**56.19**/57.43/**56.58**	**56.48**/61.60/58.10	79.56/80.70/79.92
AETHER	90.73/**93.27/91.54**	**93.61/95.03/94.06**	**98.53/98.62/98.56**	**100.00**/83.72/**94.85**	55.04/**56.50**/55.50	53.89/**63.23/56.84**	**80.33/81.55/80.71**

Table 5. **Pixel-wise Metrics of Action-Conditioned Navigation.** Comparison of performance between AETHER-no-depth and AETHER on ***in-domain/out-domain/overall*** performance. For each metric group, the better performance is highlighted in **bold**.

	PSNR ↑	SSIM ↑	MS-SSIM ↑	LPIPS ↓
AETHER-no-depth	19.13/18.67/18.97	0.5630/0.4830/0.5353	0.5467/0.5204/0.5376	0.3116/0.2995/0.3074
AETHER	**19.87/19.37/19.70**	**0.5803/0.5058/0.5545**	**0.5830/0.5627/0.5760**	**0.2691/0.2599/0.2659**

Table 6. **Quantitative Results of Action-Free Visual Path Planning.** Comparison of performance between Aether and Aether-no-depth on ***in-domain/out-domain/overall*** performance. For each metric group, the better performance is highlighted in **bold**.

	subject consistency	b.g. consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	weighted average
Aether-no-depth	88.68/89.61/88.61	93.62/93.92/93.66	98.37/98.31/98.32	**97.06**/91.67/**96.15**	54.12/56.26/54.78	51.77/58.46/54.29	79.11/80.43/79.59
Aether (Ours)	**89.69/91.61/90.36**	**93.88/94.58/94.13**	**98.50/98.40/98.46**	**97.06/91.67**/95.19	**55.83/56.87/56.19**	**54.71/61.13/56.93**	**80.21/81.53/80.67**

vancements demonstrate the evolution of world models toward dynamic, interactive, and controllable applications in robotics, gaming, and simulation.

Reconstruction. Reconstruction has been a long-standing topic in computer vision, with notable progress in both traditional and learning-based methods. Classical approaches, such as Structure-from-Motion (SfM) [9, 26, 45, 53] and Multi-View Stereo (MVS) [19, 54], rely on multi-view geometry for feature matching, pose estimation, and dense point cloud generation, demonstrating robust performance in controlled settings. Deep learning has introduced powerful alternatives, tackling sub-tasks like feature matching [14, 52], point tracking [13, 64], triangulation [42], and MVS [78, 81]. End-to-end methods now directly predict point maps [37, 66] or depth maps from images [4, 76], often incorporating camera parameters [70]. Recently, diffusion models have achieved breakthroughs in image and video generation [27, 35, 43, 73, 77], inspiring novel 3D reconstruction approaches that leverage rich 2D priors [18, 29, 34, 41, 74, 75, 88, 90]. These methods demonstrate the potential of integrating diffusion-based 2D knowledge into 3D modeling.

Video Generation. Video generation has evolved from foundational techniques like DDPM [27, 43] to modern frameworks leveraging diffusion-based techniques. Advances such as latent diffusion [51] and diffusion transformers [47] have improved generation quality, while models like Sora [5] and Stable Video Diffusion (SVD) [3] emphasize temporal consistency. Open-source models, including LTX Video [24], CogVideoX [77], and Hunyuan Video [35], offer increased flexibility, and techniques like multi-scale architectures (e.g., Pyramid Flow [32]) enhance motion dynamics. These advancements highlight rapid progress, with ongoing efforts to improve scalability and temporal stability.

7. Conclusion and Limitations

In this work, we introduce AETHER, a geometry-aware multi-task world model that reconstructs 4D dynamic videos, predicts future frames conditioned on observation images and actions, and performs visual planning based on observation and goal images. We propose an automatic 4D synthetic data labeling pipeline, enabling AETHER to train on synthetic data and generalize to unseen real-world data in a zero-shot manner. Post-trained on the CogVideoX base model, AETHER achieves state-of-the-art or competitive reconstruction performance and outperforms baselines in generation and planning tasks, demonstrating the value of incorporating reconstruction objectives into world modeling.

However, limitations remain. Camera pose estimation is less accurate, likely due to incompatibilities between raymap representation and prior video diffusion models. Indoor scene reconstruction also lags behind outdoor performance, likely due to the predominance of outdoor training data. Additionally, predictions without language prompts often fail in highly dynamic scenes. Future work can address these by exploring novel action representations, co-training with real-world data, and retaining the language prompting capabilities of the base model.

References

[1] Sameer Agarwal, Keir Mierle, et al. Ceres solver: Tutorial & reference. *Google Inc*, 2(72):8, 2012. 3

[2] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2025. 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 8

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 8

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 8

[6] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 6, 7

[7] Junyi Chen, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Where am i and what will i see: An auto-regressive model for spatial localization and view prediction, 2024. 4

[8] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7063–7072, 2019. 3

[9] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1212–1221, 2017. 8

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[11] Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a Transformer. 2024. 2

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 3

[13] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 8

[14] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 8

[15] Tom Erez, Yuval Tassa, and Emanuel Todorov. Infinite-horizon model predictive control for periodic tasks with contacts. *Robotics: Science and systems VII*, 73, 2012. 2

[16] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023. 2

[17] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024. 2

[18] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 8

[19] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 8

[20] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 7

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 6

[22] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories, 2024. 2, 7

[23] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 7

[24] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 8

[25] Annika Hagemann, Moritz Knorr, and Christoph Stiller. Deep geometry-aware camera self-calibration from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3438–3448, 2023. 3

[26] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 8

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4, 8

[28] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[29] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 5, 6, 8

[30] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7, 8

[31] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[32] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 2, 8

[33] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 3

[34] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 4, 8

[35] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Junkun Yuan, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yanxin Long, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2024. 2, 8

[36] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 6

[37] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2, 5, 6, 8

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[39] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3

[40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 4

[41] Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. Matrix3d: Large photogrammetry model all-in-one. *arXiv preprint arXiv:2502.07685*, 2025. 8

[42] Dror Moran, Hodaya Koslowsky, Yoni Kasten, Haggai Maron, Meirav Galun, and Ronen Basri. Deep permutation equivariant structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5976–5986, 2021. 8

[43] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 8

[44] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019. 6

[45] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, pages 58–77. Springer, 2024. 8

[46] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. 2, 7

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 8

[48] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya,

Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. 1

[49] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 5

[50] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 8

[52] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 8

[53] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 8

[54] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 8

[55] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024. 5, 6

[56] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 5

[57] Ziyang Song, Zerong Wang, Bo Li, Hao Zhang, Ruijie Zhu, Li Liu, Peng-Tao Jiang, and Tianzhu Zhang. Depthmaster: Taming diffusion models for monocular depth estimation. *arXiv preprint arXiv:2501.02576*, 2025. 4

[58] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 6, 7

[59] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[60] Wan Team. Wan: Open and advanced large-scale video generative models. 2025. 2

[61] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3

[62] Veo-Team, :, Agrim Gupta, Ali Razavi, Andeep Toor, Ankush Gupta, Dumitru Erhan, Eleni Shaw, Eric Lau, Frank Belletti, Gabe Barth-Maron, Gregory Shaw, Hakan Erdogan, Hakim Sidahmed, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jeff Donahue, José Lezama, Kory Mathewson, Kurtis David, Matthieu Kim Lorrain, Marc van Zee, Medhini Narasimhan, Miaosen Wang, Mohammad Babaeizadeh, Nelly Papalampidi, Nick Pezzotti, Nilpa Jha, Parker Barnes, Pieter-Jan Kindermans, Rachel Hornung, Ruben Villegas, Ryan Poplin, Salah Zaiem, Sander Dieleman, Sayna Ebrahimi, Scott Wisdom, Serena Zhang, Shlomi Fruchter, Signe Nørly, Weizhe Hua, Xinchen Yan, Yuqing Du, and Yutian Chen. Veo 2. 2024. 1

[63] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2, 5, 6

[64] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 6, 8

[65] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 2, 5, 6

[66] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 5, 6, 8

[67] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. Ieee, 2003. 5, 7

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[69] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. 2023. 2

[70] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 230–247. Springer, 2020. 8

[71] Greg Welch, Gary Bishop, et al. An introduction to the kalman filter. 1995. 7

[72] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024. 7

[73] Wanghan Xu, Xiaoyu Yue, Zidong Wang, Yao Teng, Wenlong Zhang, Xihui Liu, Luping Zhou, Wanli Ouyang, and Lei Bai. Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359*, 2025. 8

[74] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024. 2, 4, 6, 8

[75] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15238–15250, 2024. 8

[76] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 8

[77] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 4, 5, 7, 8

[78] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 8

[79] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos, 2025. 2

[80] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 6

[81] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1):199–214, 2023. 8

[82] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 2, 5, 6

[83] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[84] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance, 2024. 2

[85] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 6

[86] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision*, pages 523–542. Springer, 2022. 2, 6

[87] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 5

[88] Haoyi Zhu, Honghui Yang, Xiaoyang Wu, Di Huang, Sha Zhang, Xianglong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, Tong He, et al. Ponderv2: Pave the way for 3d foundation model with a universal pre-training paradigm. *arXiv preprint arXiv:2310.08586*, 2023. 8

[89] Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. *Advances in Neural Information Processing Systems*, 37: 77799–77830, 2024. 2

[90] Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. Spa: 3d spatial-awareness enables effective embodied representation. *arXiv preprint arXiv:2410.08208*, 2024. 2, 8