AnyView: Synthesizing Any Novel View in Dynamic Scenes

Basile Van Hoorick[1], Dian Chen[1], Shun Iwase[1], Pavel Tokmakov[1], Muhammad Zubair Irshad[1], Igor Vasiljevic[1], Swati Gupta[1], Fangzhou Cheng[1,2], Sergey Zakharov[1], and Vitor Campagnolo Guizilini[1]

[1] Toyota Research Institute, CA, USA
[2] Amazon Web Services, CA, USA

tri-ml.github.io/AnyView

Fig. 1: Enabling consistent extreme monocular dynamic view synthesis: We introduce *AnyView*, a diffusion framework that can generate videos of dynamic scenes from *any* chosen perspective, conditioned on a single input video. Our model operates end-to-end, without explicit scene reconstruction or expensive test-time optimization techniques. Existing methods tend to fail to extrapolate, largely copying the input view. More recent baselines can recover the overall structure in some cases (1st, 2nd rows), but fail when the camera trajectories become more complex (3rd row). Meanwhile, our method preserves scene geometry, appearance, and dynamics, despite working with drastically different target poses and highly "incomplete" visual observations. (D) indicates a baseline that relies on reprojected point clouds from estimated depth maps.

Abstract. Modern generative video models excel at producing convincing, high-quality outputs, but struggle to maintain multi-view and spatiotemporal consistency in highly dynamic real-world environments. In this work, we introduce **AnyView**, a diffusion-based video generation framework for *dynamic view synthesis* with minimal inductive biases or geometric assumptions. We leverage multiple data sources with various levels of supervision, including monocular (2D), multi-view static

(3D) and multi-view dynamic (4D) datasets, to train a generalist spatiotemporal implicit representation capable of producing zero-shot novel videos from arbitrary camera locations and trajectories. We evaluate AnyView on standard benchmarks, showing competitive results with the current state of the art, and propose **AnyViewBench**, a challenging new benchmark tailored towards *extreme* dynamic view synthesis in diverse real-world scenarios. In this more dramatic setting, we find that most baselines drastically degrade in performance, as they require significant overlap between viewpoints, while AnyView maintains the ability to produce realistic, plausible, and spatiotemporally consistent videos when prompted from *any* viewpoint.

1 Introduction

Generating a new video from an arbitrary camera perspective while the scene is in motion is a highly ambitious and fundamentally under-constrained task. A single input view only depicts a fraction of the world; the rest is occluded, transient, or simply unknown. New moving objects may enter the scene at any moment, and unobserved regions might be dynamic themselves, further introducing uncertainty into the generative process. Exact 4D reconstruction from such signals is therefore impractical in the general case. For many downstream uses of 4D video representations [16, 26, 47], — such as robotics, world models, simulation, telepresence, VR/AR, autonomous driving — what matters is not an exact correspondence with ground truth, but rather whether the resulting representation is realistic, temporally stable, and self-consistent across large viewpoint changes. A common problem with learned visuomotor policies, for example, is that they often suffer from brittleness under shifting camera poses [9, 32, 41, 53].

Humans routinely engage this problem in a way that is both rooted in intuition and very useful in practice: as we observe the physical world, we mentally "re-project" scenes, inferring likely layouts, object shapes, scene completions, and plausible dynamics from limited information [5, 7, 10, 21, 24, 28, 33, 38]. This is not simply a low-level reconstruction capability: it is essentially a powerful prior over shapes, semantics, materials, and motion that yields predictions that are largely viewpoint-invariant. The goal of this paper is to take a step towards solving that objective: we target perceptually realistic 4D video synthesis under extreme camera trajectories and displacements. To that end, we endow video generative models with the same inductive bias: to produce reasonable scene completions, based on a single input video, that respect scene geometry, physics, and object permanence, even when there is little overlap with the conditioning view.

Most existing *dynamic view synthesis* (DVS) approaches and benchmarks are not built for this regime [11, 12, 20, 46, 49], as they typically operate in *narrow* settings: the input and target cameras are spatially nearby, looking in similar directions, and thus methods are designed to maximize pixel metrics under limited motion, ignoring the rest of the scene. In particular, most current state of

the art DVS methods [8, 56, 59] rely on explicit 3D reconstructions (i.e., depth reprojection + image inpainting), costly test-time optimization and finetuning techniques, and support a limited set of camera trajectories.

To move away from this simplified setting, we first present **AnyView**, a novel diffusion-based DVS architecture for high-fidelity video-to-video synthesis under dramatic camera trajectory changes, capable of producing perceptually plausible and semantically consistent videos from arbitrary novel viewpoints. Our framework is purposefully light on explicit inductive biases: camera parameters are provided via dense ray-space conditioning, allowing us to support any model (including non-pinhole), and the network learns to synthesize unobserved content implicitly, guided by large-scale, diverse training data. To reach this level of implicit 4D understanding, we leverage existing video foundation models as a source of rich internet-scale 2D appearance and motion priors, and augment them by incorporating multi-view geometry and camera controllability, learned using 12 multi-domain 3D and 4D datasets.

Secondly, due to the aforementioned shortcomings of existing evaluation procedures, we assemble **AnyViewBench**, a novel benchmark that formalizes and standardizes the *extreme* DVS task across various domains (driving, robotics, and human activity), camera rigs (ego-centric and exo-centric), and camera motion patterns (fixed, linear, or complex, sometimes with changing intrinsics). Each scene provides at least two time-synchronized views, enabling rigorous metric evaluations with ground truth videos without resorting to proxy setups.

Video results are best viewed on our project webpage: tri-ml.github.io/AnyView.

2 Related Work

2.1 Video Generative Models

In recent years, significant advances have been made in video generation, leading to the development of increasingly capable generative models. Stability AI's SVD [6] pioneered video diffusion by adding temporal layers to a pre-trained image diffusion network [37], allowing coherent short video clip generation from single images or text prompts. CogVideoX [55] introduced a 3D Variational Autoencoder (VAE) to compress videos across spatial and temporal dimensions, enhancing both compression rate and video fidelity. NVIDIA's Cosmos [2] introduced a suite of models with strong long-range temporal consistency and flexible conditioning signals (text, image and video input). Wan [45] is a novel mixture of experts-based video generation architecture, and provides a suite of video world models that excel at prompt following and photorealistic generation. However, none of these architectures were originally designed with camera conditioning in mind, focusing instead on future frame forecasting in the single – or more recently multi-camera [30] – setting.

2.2 Dynamic View Synthesis

Dynamic view synthesis is the task of generating novel renderings from arbitrary viewpoints and timesteps given a monocular video of a dynamic scene. A

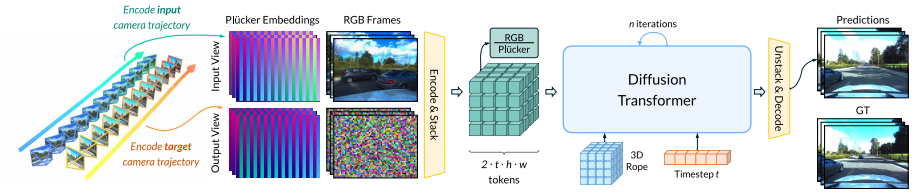


Fig. 2: The AnyView architecture. For both the clean input and noisy target videos, we concatenate pixels (RGB values) and camera information (Plücker vectors) belonging to the *same* viewpoint along the *channel* dimension, after independently encoding each modality into latent embeddings. We then stack these two multimodal videos along the *sequence dimension*, for a total of $2 \cdot t \cdot h \cdot w$ tokens, which are fed into the diffusion transformer to iteratively denoise the target video.

number of works have combined video generation with explicit geometric conditioning to improve geometric 3D consistency and control [17, 44, 50, 61]. Shape of Motion [46] addresses monocular dynamic reconstruction by representing scene motion through a compact set of SE(3) motion bases, enabling soft segmentation into multiple rigidly moving parts using monocular depth and long-range 2D tracks. It fuses monocular depth and long-range 2D tracks to obtain a globally consistent dynamic 3D representation.

While explicit modeling approaches can achieve relatively high accuracy, they are computationally expensive and brittle. GCD [43] proposed to address dynamic view synthesis as an implicit problem, by re-purposing internet-scale video diffusion models via camera conditioning. This implicit formulation provides the greatest flexibility and robustness, but requires ground truth multi-view video data for training. ReCamMaster [4] advanced this research direction by utilizing a more powerful video generation model and a more realistic simulator to generate training data, whereas Trajectory Attention [51] augments video diffusion models with a trajectory-aware attention mechanism, improving fine-grained camera motion control and temporal consistency. AC3D [3] analyzes how video diffusion models internally represent 3D camera motion, adding ControlNet-style conditioning to improve controllability.

Other methods [8, 36, 59] have taken a hybrid approach by first lifting the input video in 3D via monocular depth estimation, reprojecting the resulting point cloud to the target camera pose, and then treating dynamic view synthesis as an inpainting problem. Among these methods, CogNVS [8] further introduces test-time optimization to improve rendering accuracy at the cost of inference speed, while StreetCrafter [52] focuses on autonomous driving scene generation, utilizing LiDAR renderings as the control signal. Very recently, InverseDVS [57] has proposed a training-free approach that reformulates inpainting as structured latent manipulation in the noise initialization phase of a video diffusion model.

While the shift towards explicit scene reconstruction and test-time optimization has led to high-quality dynamic view synthesis in the *narrow* setting, where camera motion is limited to neighboring and highly overlapping regions, we ex-

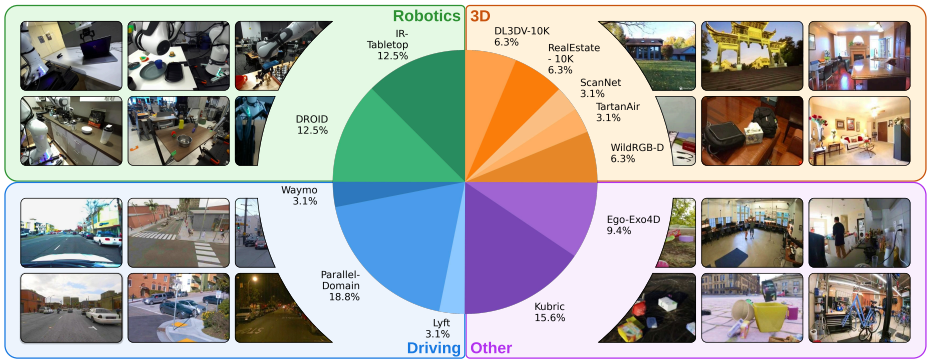


Fig. 3: Overview of our training data mixture. We train and evaluate AnyView on both single-view and multi-view videos from four domains: *3D*, *Driving*, *Robotics*, and *Other* (see Section 3.3). During training, we perform weighted sampling to ensure each domain is seen equally often, *i.e.* comprises 25% of the batch.

perimentally demonstrate that these methods do not generalize to the more challenging *extreme* setting. In contrast, data-driven, implicit approaches are in principle capable of dynamic view synthesis from any viewpoint, but are in practice limited by the availability of diverse training data. In this work, we address this by (1) combining a wide body of publicly available datasets to train AnyView — the first model capable of synthesizing arbitrary novel views in dynamic, real-world scenes; and (2) proposing a new benchmark, AnyViewBench, to properly evaluate dynamic view synthesis performance in this new setting.

3 Methodology

3.1 Problem Statement

The goal of dynamic view synthesis (DVS) is to create an output video $\boldsymbol{V}_y$ of an underlying scene as depicted from a chosen virtual viewpoint c_y, given an input video $\boldsymbol{V}_x$ recorded by a camera with known poses c_x and intrinsics i_x over time. Specifically, we define the input (observed) RGB video as $\boldsymbol{V}_x \in \mathbb{R}^{T\times H\times W\times 3}$, the target (unobserved) RGB video as $\boldsymbol{V}_y \in \mathbb{R}^{T\times H\times W\times 3}$, the input camera trajectory as $c_x \in \mathbb{R}^{T\times 4\times 4}$ with intrinsics $i_x \in \mathbb{R}^{T\times 3\times 3}$, and the target camera trajectory as $c_y \in \mathbb{R}^{T\times 4\times 4}$ with intrinsics $i_y \in \mathbb{R}^{T\times 3\times 3}$. Using a generative model f, we estimate $\boldsymbol{V}_y$ corresponding to the desired novel viewpoint c_y by drawing from a conditional probability distribution:

$$\boldsymbol{V}_y \sim P_f\left(\boldsymbol{V}_y \mid \boldsymbol{V}_x, c_x, i_x, c_y, i_y\right) \tag{1}$$

The camera parameters c_x, i_x, c_y, i_y represent two sequences of fully specified 6-DoF $SE(3)$ camera poses, ensuring that the task setting is both general and unambiguous. Moreover, there should be some spatial overlap in content between the two perspectives c_x and c_y (even if this overlap is temporally asynchronous), otherwise the conditioning signal loses its relevance.

3.2 Architecture

The task described above involves (1) *synthesis of high-dimensional data* in the form of multiple images, and (2) *considerable uncertainty handling* mainly because of occlusion and ambiguous object motion. These requirements are challenging, but naturally lend themselves to being implemented using the generative video paradigm. Hence, we adopted Cosmos [30], a latent diffusion transformer, as our underlying base representation, for its efficiency, high-quality pretrained checkpoints, and flexible conditioning mechanisms (*e.g.* text, image, and video).

Our proposed AnyView architecture, illustrated in Figure 2, prioritizes simplicity and scalability. Contrary to most state-of-the-art methods [8, 36, 51, 59], we elect to not use warped depth maps as explicit conditioning, owing to the risk of compounding errors from depth estimation, and instead rely solely on a learned implicit representation as our rendering mechanism. The reasoning behind this decision is so that we can achieve *unbounded* dynamic view synthesis, that does not require substantial overlap between target and generated videos, thus allowing for more extreme camera motion. We explore this property in our proposed AnyViewBench, outperforming baselines that rely on explicit reprojection.

In order to make AnyView 4D-aware and controllable, we feed information about both viewpoints into the network in a structured yet straightforward way. To account for the possible lack of an absolute frame of reference, all camera poses are expected to exist relative to the target viewpoint $c_{y,0}$ at time $t = 0$. In other words, c_y always starts at the "origin", with $c_{y,0} = I_{4\times4}$, mapping to the identity matrix. If this is not the case, a simple change of coordinate system can be done by applying $\tilde{c} = c \cdot c_{y,0}^{-1}$, assuming the camera-to-world extrinsics convention.

First, the given video $\boldsymbol{V}_x$ is compressed into a latent space by a video tokenizer to become $\boldsymbol{v}_x \in \mathbb{R}^{t\times h\times w\times d}$, with spatiotemporal downsampling ratios $T/t = 4$ and $H/h = W/w = 8$, and embedding size $d = 16$. We then encode all camera parameters c_x, i_x, c_y, i_y into a unified *Plücker representation* $\boldsymbol{P} = (\boldsymbol{r}, \boldsymbol{m})$ [18], which combines extrinsics and intrinsics into a dense map containing per-pixel *ray* vectors $\boldsymbol{r}$ and *moment* vectors $\boldsymbol{m} = \boldsymbol{r} \times \boldsymbol{o}$. This results in two quantities $\boldsymbol{P}_x, \boldsymbol{P}_y \in \mathbb{R}^{T\times H\times W\times 6}$, *i.e.* tensors with the same dimensionality as a 6-channel video, or two 3-channel videos. We can therefore separately tokenize the rays $\boldsymbol{r} \in \mathbb{R}^{T\times H\times W\times 3}$ and moments $\boldsymbol{m} \in \mathbb{R}^{T\times H\times W\times 3}$ (shown as alternating columns in Figure 2) the same way as before into $\boldsymbol{p}_x, \boldsymbol{p}_y \in \mathbb{R}^{t\times h\times w\times 2d}$. An interesting property of using Plücker maps instead of direct camera conditioning [43] is the natural handling of non-pinhole camera models, since the 3D ray vectors directly capture camera intrinsics in a general, non-parametric way.

Because latent RGB and Plücker tokens from each viewpoint contain information pertaining to the same spatiotemporal region, we merge them via concatenation along the channel dimension, while keeping tokens from separate viewpoints separate. Since there are two viewpoints in total, this results in a sequence of $2 \cdot t \cdot h \cdot w$ tokens, each of length $3 \cdot d$. All tokens are tagged with rotary positional embeddings [39], as well as a unique per-view embedding. After

Table 1: Testing datasets. We evaluate on several benchmarks that cover both *narrow* and *extreme* settings. We define **AnyViewBench** as a multi-faceted benchmark focusing on the latter category, setting a new standard for consistent dynamic view synthesis in challenging settings. Test splits are capped at 64 per dataset with uniform subsampling. *Input Cam.* refers to what the camera characteristic of the observed video (*i.e.* static vs dynamic). The rightmost column (*Generalization Type*) qualitatively denotes how large the distribution shift is relative to the AnyView training mixture.

Benchmark	S/R	Domain	Type	#Cameras	#Episodes	Resolution	Input Cam.	Gen. Type
Narrow dynamic view synthesis								
DyCheck iPhone [12]	Real	Hand-Object	4D	2 – 3	5 / 7	288 × 384	Moving	0-shot overall
Kubric-4D (gradual) [14]	Sim	Multi-Object	4D	16	20 / 100	384 × 256	Fixed	In-dist.
ParDom-4D (gradual) [1]	Sim	Driving	4D	19	20 / 61	384 × 256	Variable	In-dist.
***AnyViewBench*: In-distribution extreme** dynamic view synthesis								
DROID (ID) [23]	Real	Robotics	4D	2	64 / 3,301	384 × 208	Fixed	In-dist.
Ego-Exo4D (ID) [13]	Real	Human Activity	4D	4 – 5	64 / 276	384 × 208	Fixed	In-dist.
LBM	Sim + Real	Robotics	4D	2	64 / 5,988	336 × 256	Fixed	In-dist.
Kubric-4D (direct) [14]	Sim	Multi-Object	4D	16	20 / 100	384 × 256	Fixed	In-dist.
Kubric-5D [14]	Sim	Multi-Object	4D	16	64 / 200	384 × 256	Variable	In-dist.
Lyft [19]	Real	Driving	4D	6	64 / 436	384 × 320	Variable	In-dist.
ParDom-4D (direct) [1]	Sim	Driving	4D	19	20 / 61	384 × 256	Variable	In-dist.
Waymo [40]	Real	Driving	4D	5	64 / 202	384 × {176, 256}	Variable	In-dist.
***AnyViewBench*: Zero-shot extreme** dynamic view synthesis								
Argoverse [48]	Real	Driving	4D	7	64 / 1,042	{288, 384} × {288, 384}	Variable	dataset
AssemblyHands [31]	Real	Hand-Object	4D	8	20 / 20	384 × 208	Fixed	domain
DDAD [15]	Real	Driving	4D	6	64 / 200	384 × 240	Variable	dataset
DROID (OOD) [23]	Real	Robotics	4D	2	64 / 252	384 × 208	Fixed	station
Ego-Exo4D (OOD) [13]	Real	Human Activity	4D	4 – 5	64 / 408	384 × 208	Fixed	activity/site

completing all self-attention and cross-attention blocks, the output sequence is the latent RGB video $\boldsymbol{v}_y \in \mathbb{R}^{t \times h \times w \times d}$. During training, these latent tokens are supervised with an $\mathcal{L}_2$ loss, and during inference, they are iteratively denoised before finally being decoded into a generated video $\boldsymbol{V}_y \in \mathbb{R}^{T \times H \times W \times 3}$.

3.3 Datasets

Because AnyView does not rely on any explicit conditioning mechanism (*e.g.* intermediate depth maps) to facilitate the rendering of novel viewpoints, it must learn implicit multi-view geometry as well as a wide range of appearance priors, to be able to inpaint and outpaint potentially large unobserved portions of the scene. In order to train such a generalist spatiotemporal representation capable of handling multiple domains, we combined 12 different 4D datasets into our unified training pipeline. Among them is *Kubric-5D*, our newly introduced variation of Kubric-4D [14,43] that vastly increases the diversity of camera trajectories. We classify our training datasets into four distinct quadrants: *Robotics*, *Driving*, *3D*, and *Other*. A visual overview is illustrated in Figure 3, and more details are provided in the supplementary material. To the best of our knowledge, this data mixture covers a significant portion of publicly available multi-view video datasets. We leave the inclusion of additional 4D data [34,35,62] to future work.

3.4 Implementation Details

We train AnyView for 40,000 iterations on 64 NVIDIA H200 GPUs at a global batch size of 512. We apply curriculum learning with increasing resolution: first

we train at a largest image dimension of 384 for 30,000 steps, before finetuning at a largest image dimension of 576. The initial learning rate is $5 \cdot 10^{-5}$, and drops smoothly to $1 \cdot 10^{-5}$ according to a cosine schedule. All experiments are performed with the `Cosmos-Predict2-2B-Video2World` [29] model, starting from their pretrained network, which has around 2 billion parameters. We disable language conditioning, since it is not relevant to our task setting. Furthermore, in order to properly combine datasets with varying physical scales, we divide the translation vectors of all cameras $\{c_x, c_y\}$ by a carefully chosen per-dataset normalization constant to ensure the resulting Plücker values always fall in the range $[-1, 1]$, occasionally clipping pixels as needed.

4 Experiments

4.1 Evaluation Challenges

As the field is evolving, many existing DVS benchmarks are beginning to **lack difficulty**, containing scenes with minimal object motion and modest camera transformations [12, 27, 43, 58]. Qualitative results are often demonstrated on camera trajectories with rotational variations of only about $10 - 30$ degrees relative to the center of the scene [4, 43, 46, 56, 60]. Consequently, the heavy lifting of inpainting large occlusions is mostly avoided, making it unclear to which extent these models learn robust, multi-view consistent 4D representations. These efforts are further complicated by a **lack of standardization**, which can be partially attributed to the inherent complexity of DVS: merely describing the task is insufficient to define a path towards practical execution. Design choices often left in the dark include but are not limited to: video resolution, number of frames, camera controllability and conventions, used frames of reference, the space of possible camera transformations, and so on.

4.2 Benchmarks

We first consider three popular DVS benchmarks that can be classified as falling into the "narrow" regime. Then, to address the aforementioned concerns, we propose *AnyViewBench*, which substantially pushes models into the more challenging "extreme" regime.

DyCheck iPhone (narrow DVS). The iPhone dataset [12] is a small collection of high-quality, real-world, multi-view videos of easy-to-moderate difficulty established to measure DVS fidelity. Following previous works [8], that have pointed out that the provided camera poses are not very accurate, we compute corrected extrinsics using MoSca [25].

Kubric-4D and ParDom-4D (narrow + extreme DVS). The GCD [43] paper introduced two synthetic datasets for DVS training and evaluation, based on the Kubric [14] and ParallelDomain [1] simulation environments.

AnyViewBench (extreme DVS). We introduce AnyViewBench, a multi-faceted benchmark that covers datasets across multiple domains (driving, robotics,

Table 2: Narrow DVS results. We compare against several state-of-the-art baselines, including those using test-time optimization (TTO) and auxiliary networks (Aux.) for depth (D), poses (P), and/or 2D point tracks (T). The inference runtime assumes that a video was not observed before, and thus includes a test-time optimization stage if present. Results reported by: †original paper; ‡another paper (cited); *computed by us. Gray rows are ablations of AnyView.

Method	PSNR↑	SSIM↑	LPIPS↓	TTO	Aux.	0-shot	Time
DyCheck iPhone [12]							
ShapeOfMotion† [46]	16.72	0.630	0.450	✓	D P T	–	~1h
CogNVS† [8]	16.94	0.449	0.598	✓	D(GT)	–	~1h
GEN3C* [36]	10.13	0.175	0.695		D	✓	~15m
TrajAttn* [51]	10.30	0.181	0.682		D	✓	~10m
TrajCrafter† [59]	14.24	0.417	0.519		D	✓	~10s
GCD* [43]	11.43	0.247	0.728		–	✓	~10s
AnyView	13.47	0.295	0.550		–	✓	~10s
w/ GCD data	11.38	0.191	0.662		–	✓	~10s
from scratch	9.40	0.173	0.732		–	✓	~10s
Kubric-4D (gradual) [14]							
CogNVS† [8]	22.63	0.760	0.232	✓	D(GT)	–	~1h
ReCapture† [60]	20.92	0.596	0.402	✓	D	–	~15m
GEN3C† [36]	19.41	0.630	0.290		D	✓	~15m
TrajAttn* [51]	15.73	0.404	0.530		D	✓	~5m
TrajCrafter‡ [8]	20.93	0.730	0.257		D	✓	~10s
GCD* [43]	20.42	0.581	0.405		–		~10s
AnyView	21.21	0.644	0.358		–		~10s
w/ GCD data	22.06	0.611	0.343		–		~10s
from scratch	17.68	0.445	0.499		–		~10s
ParDom-4D (gradual) [1]							
CogNVS† [8]	24.34	0.797	0.302	✓	D(GT)	–	~1h
GEN3C* [36]	18.40	0.528	0.542		D	✓	~15m
TrajAttn* [51]	20.03	0.566	0.518		D	✓	~5m
TrajCrafter‡ [8]	21.46	0.719	0.342		D	✓	~10s
GCD* [43]	24.75	0.724	0.355		–		~10s
AnyView	26.29	0.758	0.320		–		~10s
w/ GCD data	26.18	0.748	0.345		–		~10s
from scratch	21.89	0.590	0.507		–		~10s

and human activities), as shown in Table 1. The camera motion patterns range from simple (fixed or linear) to complex (*e.g.* highly non-linear trajectories, changing intrinsics, *etc.*). To promote rigorous evaluation, we provide synchronized videos from at least two separate viewpoints for each episode, with well-defined details such as spatial resolution and number of frames, such that ground truth metrics can be calculated in a straightforward manner.

For all *in-distribution* datasets, we separate roughly 10% to serve as validation, and for both *in-distribution* and *zero-shot* datasets, we curate smaller

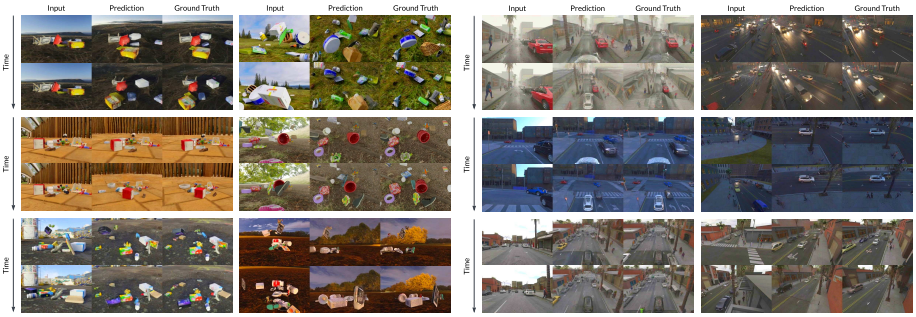


Fig. 4: AnyView in-domain DVS results on **Kubric-4D** (left) and **Pardom-4D** (right). We show the first and last frame of each video. The scene layout is generally preserved very well, despite drastic viewpoint changes and/or heavy occlusion from the input vantage point.

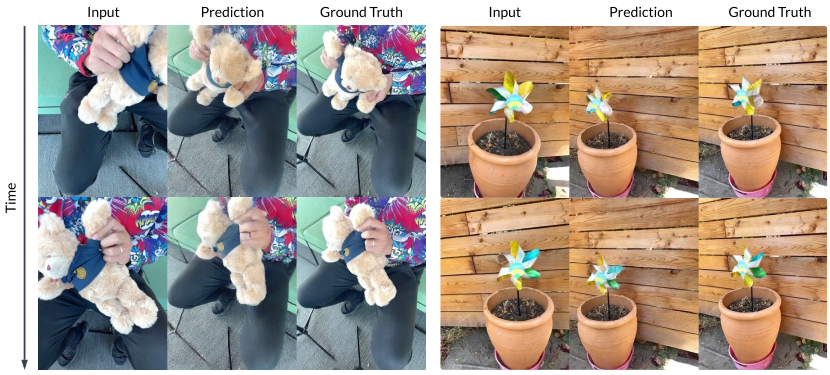


Fig. 5: Results on DyCheck iPhone (0-shot narrow DVS). While not highly dynamic, these scenes contain subtle, intricate motions and hand-object interactions.

subsets to serve as official test splits. Moreover, two DROID stations (GuptaLab, ILIAD), as well as certain EgoExo4D institutions (FAIR, NUS) and activities (CPR, Guitar), are held out to serve as *zero-shot* evaluation. More information about AnyViewBench can be found in the supplementary material, and we will release it upon publication.

4.3 Baselines

Most current DVS methods face key limitations: the input video must be captured from a strictly *static* camera [43], or from a strictly *dynamic* camera [8,36], or both input and output videos must start from the same position [4,51,59], or the camera controlling mechanism has limited degrees of freedom [4,43,59]. As a result, methods that excel in certain conditions might be incompatible with slightly different evaluation settings, hindering standardized evaluation across multiple benchmarks. More information detailing all prior works we considered

Table 3: Extreme DVS results (direct GT comparisons). Note that *in-distribution* datasets are part of AnyView's training mixture, but might be zero-shot for some of the baselines, hence we provide these results for completeness. For qualitative comparison, please refer to Figure 1.

Dataset	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓
***AnyViewBench*: In-distribution**												
	GCD [43]				TrajAttn [51]				GEN3C [36]			
DROID [23]	10.18	0.255	0.688	5699	9.93	0.247	0.671	3430	9.62	0.211	0.666	3359
EgoExo4D [13]	12.10	0.255	0.670	3353	11.64	0.234	0.646	2520	11.69	0.222	0.643	2097
LBM	12.59	0.398	0.694	2458	13.47	0.421	0.614	1891	13.48	0.449	0.581	1943
Kubric-4D [14]	17.57	0.477	0.512	3731	13.47	0.320	0.607	5257	13.59	0.341	0.599	4556
Kubric-5D	13.83	0.391	0.644	3405	13.25	0.360	0.628	3661	13.10	0.327	0.627	3241
Lyft [19]	8.72	0.335	0.697	6442	8.33	0.273	0.621	1819	8.43	0.286	0.634	2211
ParDom-4D [1]	22.67	0.656	0.457	3579	16.91	0.445	0.610	6027	16.64	0.478	0.590	5738
Waymo [40]	12.93	0.393	0.647	4344	12.55	0.350	0.613	1617	12.98	0.350	0.600	1653
Average	13.95	0.400	0.623	4126	12.44	0.331	0.626	3278	12.44	0.333	0.617	3100
	TrajCrafter [59]				CogNVS [8]				Ours (AnyView)			
DROID [23]	10.45	0.257	0.620	3416	9.44	0.281	0.634	4468	14.47	0.445	0.472	1399
EgoExo4D [13]	11.33	0.195	0.642	3367	10.80	0.241	0.670	2607	18.14	0.531	0.379	995
LBM	13.68	0.447	0.537	2090	13.30	0.453	0.548	1903	17.94	0.649	0.348	594
Kubric-4D [14]	14.14	0.294	0.592	5357	12.55	0.329	0.601	5104	18.38	0.441	0.362	2009
Kubric-5D	13.30	0.287	0.625	3450	12.18	0.318	0.643	4563	17.18	0.468	0.428	1130
Lyft [19]	8.73	0.251	0.621	2444	8.34	0.319	0.628	3462	15.37	0.564	0.371	719
ParDom-4D [1]	18.23	0.475	0.586	5790	18.36	0.499	0.564	5668	24.26	0.688	0.351	2157
Waymo [40]	12.66	0.312	0.593	2074	13.27	0.377	0.594	2186	16.52	0.477	0.480	907
Average	12.82	0.315	0.602	3499	12.28	0.352	0.610	3745	**17.78**	**0.533**	**0.399**	**1239**
***AnyViewBench*: Zero-shot**												
	GCD [43]				TrajAttn [51]				GEN3C [36]			
Argoverse [48]	11.45	0.403	0.682	3536	10.62	0.317	0.665	1717	10.52	0.319	0.680	1620
AssemblyHands [31]	9.77	0.262	0.759	3956	9.97	0.266	0.736	2643	9.86	0.237	0.732	1742
DDAD [15]	9.81	0.278	0.660	5026	9.16	0.244	0.620	1977	9.35	0.259	0.600	1645
DROID [23]	11.81	0.315	0.690	5358	10.83	0.320	0.678	3129	10.56	0.276	0.674	3232
EgoExo4D [13]	11.98	0.239	0.668	3862	11.31	0.203	0.653	2576	11.40	0.193	0.651	2210
Average	10.96	0.299	0.692	4348	10.38	0.270	0.670	2408	10.34	0.257	0.667	2090
	TrajCrafter [59]				CogNVS [8]				Ours (AnyView)			
Argoverse [48]	10.67	0.325	0.621	1747	10.76	0.360	0.610	1469	12.38	0.399	0.587	982
AssemblyHands [31]	11.45	0.248	0.701	4300	9.93	0.281	0.701	3862	11.21	0.291	0.688	2870
DDAD [15]	10.73	0.300	0.558	2274	10.81	0.355	0.572	1879	11.44	0.341	0.519	1419
DROID [23]	11.37	0.339	0.614	3395	10.48	0.358	0.632	4312	12.34	0.422	0.601	1956
EgoExo4D [13]	11.27	0.180	0.647	3303	10.52	0.227	0.683	2967	13.30	0.297	0.562	1555
Average	11.10	0.279	0.628	3004	10.50	0.316	0.640	2898	**12.13**	**0.350**	**0.591**	**1757**

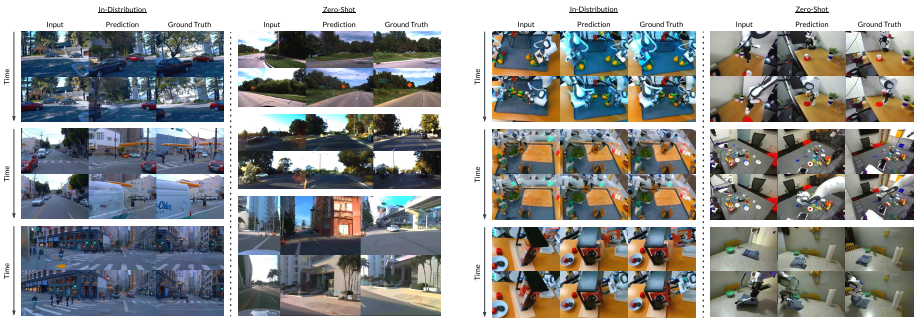


Fig. 6: AnyView extreme DVS results on **driving** (left) and **robotics** (right) benchmarks. We show both in-domain and zero-shot results. For driving videos, we focus on the three frontal cameras, whereas for robotics, we focus on all scene cameras.

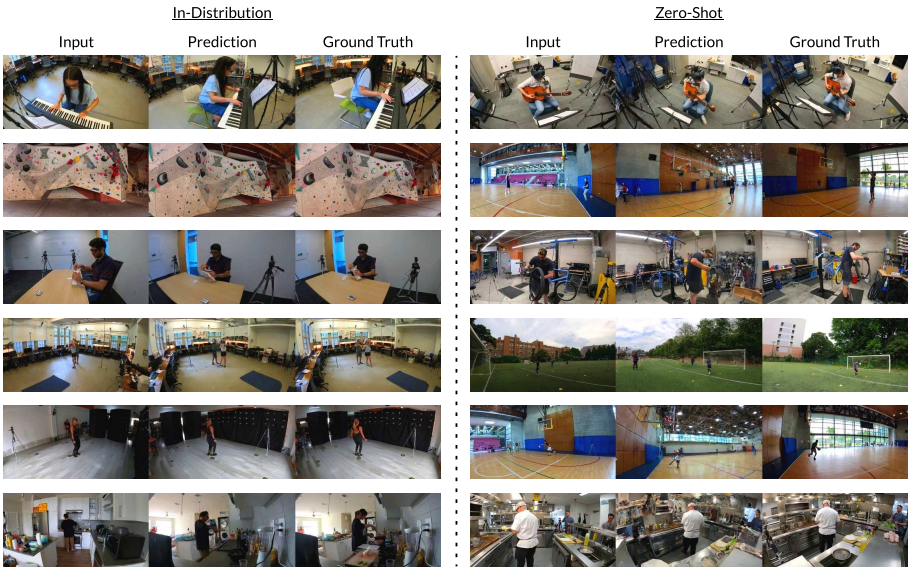


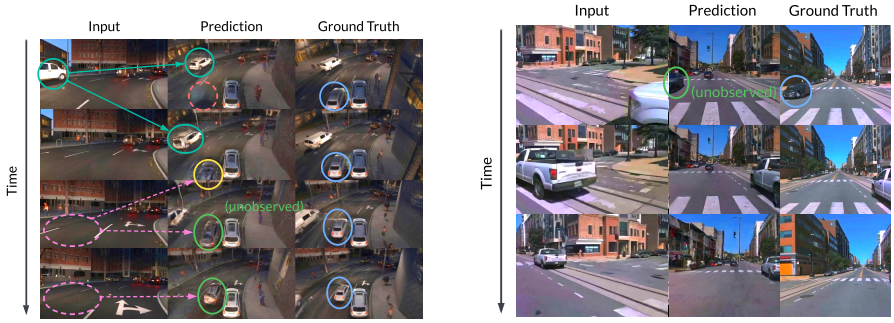
Fig. 7: AnyView extreme DVS results on Ego-Exo4D. We show both in-domain and zero-shot results. Note that in the zero-shot case, the background often has to be "guessed" from the other camera viewpoint, but the inpainted regions (see *e.g.* basketball, soccer) integrate harmoniously with the rest of the scene.

as baselines can be found in the supplementary material. Most of these models already evaluate on at least a subset of the "narrow" benchmarks, but we additionally evaluate them (doing our best effort to project down to and accommodate the space of supported camera transformations as needed) on AnyViewBench, which embodies the "extreme" benchmarks. When evaluating baseline methods that require depth estimation to render reprojected images, we use DepthAnythingV2 [54] and tune the maximum depth parameter for each dataset to achieve the best alignment between reprojected and ground truth images.

4.4 Results

Metrics. Following standard convention, we report DVS results in terms of PSNR (dB), SSIM, and LPIPS (VGG), averaged over all frames in the generated video, as well as FVD [42]. Note that pixel-level metrics can only attest to how similar generated predictions are to the ground truth, but not necessarily how realistic and plausible they are when the true underlying scene cannot be fully known due to lack of overlap between viewpoints; FVD complements these by capturing perceptual realism independently of exact pixel correspondence.

Narrow DVS. Quantitative results on existing narrow DVS benchmarks are reported in Table 2, with qualitative results in Figs. 4 and 5. For completeness, we also include metrics as reported by other papers, as well as evaluate the baselines ourselves when possible. AnyView outperforms GCD [43], the only baseline that does not require explicit depth estimation or reprojection, by a large margin, and



(a) Headlight reflections. While the white van is partially visible in the first few frames, and thus gets depicted accurately in the remainder of the video, the white car is never observed in *any* input frame. Instead, AnyView appears to pick up on the headlights reflecting on the road. Although the reconstructed car does not have the correct appearance, the model estimates its trajectory by tracking the reflection over time.

(b) Driving behavior. This zero-shot ArgoVerse scene depicts the ego vehicle pausing for a moment, and then turning left. The model correctly hallucinates the black car passing by on the left of the generated video, despite never observing it, presumably based on the suspicion that the driver must be waiting at the green light because of oncoming traffic before executing the unprotected left turn.

Fig. 8: Examples of advanced reasoning within AnyView, as a way to indirectly guide generation in unobserved parts of the scene.

compares favorably with explicit depth reprojection methods — and those that require expensive test-time-optimization — in most metrics. This *narrow* setting (i.e., large overlapping regions with small viewpoint changes) is particularly well-suited for such methods, since a lot of information can be directly transferred across viewpoints, and the model is tasked solely with inpainting missing regions.

Extreme DVS. Next, we report results in *extreme* DVS setting using AnyView-Bench, with quantitative results in Table 3 and illustrations in Figs. 6 and 7.[3] These scenarios are much more challenging, since they require implicit 4D understanding to ensure spatiotemporal consistency. For example, in real-world driving, the amount of spatial overlap between neighboring cameras is generally small, meaning that when the model is prompted with generating the front-left view based solely on the front view (or vice-versa), it has to plausibly infer the majority of the scene based on little information. However, if the ego vehicle is moving, information is able to eventually "leak" into other views and propagate across the entire sequence, further limiting the space of "correct" generations.

Geometric consistency. To assess whether AnyView's implicit representation captures 3D geometry rather than merely plausible pixels, we estimate camera extrinsics from the *generated* videos with MapAnything [22] and compare them to ground-truth estimates. As reported in Table 4, AnyView outperforms all

[3] FVD follows the official TF-GAN implementation with the I3D network; high values are due to very low sample size per dataset ($20 - 64$ instead of $\gg 1000$).

Table 4: Camera pose estimation results on AnyViewBench. *Transl.* and *Rot.* are the translation and rotation (degrees) errors between camera extrinsics estimated from the generated versus ground-truth videos; AUC_{15} is the area under the pose-accuracy curve at a binning threshold of 15.

Method	In-distribution			Zero-shot		
	Transl.↓	Rot.↓	AUC_{15}↑	Transl.↓	Rot.↓	AUC_{15}↑
GCD [43]	<u>0.919</u>	2.774	19.87	<u>1.201</u>	<u>2.381</u>	12.14
TrajAttn [51]	2.219	5.872	11.89	2.376	5.156	8.01
GEN3C [36]	1.798	5.388	13.13	2.234	4.792	7.76
TrajCrafter [59]	1.021	3.101	<u>21.30</u>	1.421	2.571	11.79
CogNVS [8]	1.215	<u>2.490</u>	17.42	1.398	2.759	<u>12.31</u>
Ours (AnyView)	**0.217**	**0.654**	**43.38**	**0.608**	**1.781**	**16.52**

Table 5: Extreme DVS ablations. We compare AnyView with two ablations of the same model: one trained from scratch (*i.e.* without Cosmos pretrained weights), and another finetuned with GCD training data only (*i.e.* Kubric-4D and ParallelDomain-4D) [43]. All ablations use the same underlying architecture.

Dataset	From scratch			GCD data			AnyView		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
***AnyViewBench*: In-distribution**									
DROID	15.36	0.462	0.464	10.53	0.253	0.687	14.47	0.445	0.472
EgoExo4D	16.24	0.400	0.500	12.36	0.252	0.686	18.14	0.531	0.379
LBM	19.07	0.676	0.317	12.64	0.372	0.652	17.94	0.649	0.348
Kubric-4D	16.17	0.377	0.547	19.77	0.506	0.410	18.38	0.441	0.362
Kubric-5D	14.57	0.373	0.577	14.83	0.370	0.608	17.18	0.468	0.428
Lyft	11.67	0.411	0.510	8.74	0.268	0.691	15.37	0.564	0.371
ParDom-4D	21.02	0.539	0.525	24.50	0.703	0.378	24.26	0.688	0.351
Waymo	13.95	0.406	0.558	12.93	0.350	0.635	16.52	0.477	0.480
Average	<u>16.01</u>	<u>0.456</u>	<u>0.500</u>	14.54	0.384	0.593	**17.78**	**0.533**	**0.399**
***AnyViewBench*: Zero-shot**									
Argoverse	11.39	0.313	0.655	11.02	0.346	0.673	12.38	0.399	0.587
AssemblyHands	10.83	0.215	0.708	10.55	0.269	0.739	11.21	0.291	0.688
DDAD	10.51	0.288	0.586	9.68	0.249	0.637	11.44	0.341	0.519
DROID	11.46	0.314	0.622	11.49	0.335	0.680	12.34	0.422	0.601
EgoExo4D	12.67	0.206	0.632	12.27	0.230	0.682	13.30	0.297	0.562
Average	<u>11.37</u>	0.267	<u>0.641</u>	11.00	<u>0.286</u>	0.682	**12.13**	**0.350**	**0.591**

baselines in translation and rotation accuracy by a large margin, in-distribution and zero-shot, despite using no explicit geometric supervision, unlike several baselines (*e.g.* GEN3C, which relies on DepthAnythingV2 [54]). This indicates a genuine, implicit understanding of 3D structure.

Ablations. We also ablate two key design choices of AnyView in Tables 2 and 5: (1) training from scratch, as opposed to initializing from Cosmos pretrained weights; and (2) match the GCD training data by only training on Kubric-4D and ParallelDomain-4D [43], as opposed to using our full AnyView data mixture as in Figure 3. Training from scratch consistently degrades performance across both narrow and extreme settings, confirming the importance of leveraging a strong

video foundation model. Using only GCD training data yields competitive results on Kubric-4D and ParallelDomain-4D, presumably because they are in-domain, but generalizes poorly to other domains, particularly on DyCheck iPhone. The full AnyView training mixture, combining diverse real-world and synthetic data, provides the best overall performance and strongest zero-shot generalization.

Analysis. Remarkably, in the upper left scenario in Figure 6, the red car arriving at the intersection is predicted on the left view *before* it is visible in the input front view, showing that AnyView has learned to maintain spatiotemporal consistency, leading to improved performance in areas that otherwise would be ill-defined. A related behavior is also observed in the left examples of Figure 7, where AnyView leverages its foundational knowledge to infer how a basketball court or soccer field should look like from different perspectives. Moreover, in Fig. 8 we show anecdotal examples of AnyView leveraging subtle visuals cues to improve generation accuracy in unobserved areas, as evidence of advanced common sense and spatiotemporal reasoning.

Implicitly learning these useful spatiotemporal properties in a data-driven way enables AnyView to produce more realistic and physically plausible representations of real-world scenarios compared to all baselines. As shown in Fig. 1, while methods that rely on potentially inaccurate depth reprojection (*e.g.* TrajAttn and GEN3C) struggle when starting from target poses away from input poses, AnyView successfully generates smooth, consistent target scenes regardless of camera positioning. Similarly, AnyView is able to accurately outpaint much larger unobserved portions of the scene compared to methods trained mostly for limited inpainting (*e.g.* TrajCrafter and CogNVS). As a consequence of these useful properties, we achieve state of the art zero-shot DVS performance on AnyViewBench, outperforming all other baseline methods by a significant margin across all considered datasets.

5 Discussion

In this paper, we propose *AnyView*, a generalist dynamic view synthesis framework targeting extreme camera displacements. We also contribute *AnyViewBench*, a well-rounded benchmark that focuses on highly challenging scenarios from various domains, showing that AnyView significantly outperforms baselines in such settings with large camera displacement and limited overlap between views. We remark that AnyView is not without limitations: it requires known input camera poses, offers no hard geometric guarantees, and can still produce plausible yet incorrect content in heavily occluded regions, particularly under the most extreme viewpoint changes. Regardless, we hope that this work provides a useful building block towards improving video foundation models and 4D representations, with potential applications in dynamic scene reconstruction, world models, robotics, self-driving, and more.

References

1. Parallel domain. https://paralleldomain.com/ (2024) 7, 8, 9, 11

2. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025), https://arxiv.org/abs/2501.03575 3
3. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22875–22889 (2025) 4
4. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647 (2025), https://arxiv.org/abs/2503.11647 4, 8, 10
5. Baillargeon, R.: Object permanence in 31/2-and 41/2-month-old infants. Developmental psychology **23**(5), 655 (1987) 2
6. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023), https://arxiv.org/abs/2311.15127 3
7. Burgess, N.: Spatial memory: how egocentric and allocentric combine. Trends in cognitive sciences **10**(12), 551–557 (2006) 2
8. Chen, K., Khurana, T., Ramanan, D.: Reconstruct, inpaint, finetune: Dynamic novel-view synthesis from monocular videos (2025), https://arxiv.org/abs/2507.12646 3, 4, 6, 8, 9, 10, 11, 14
9. Chen, L.Y., Xu, C., Dharmarajan, K., Irshad, Z., Cheng, R., Keutzer, K., Tomizuka, M., Vuong, Q., Goldberg, K.: Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In: Conference on Robot Learning (CoRL). Munich, Germany (2024) 2
10. DiCarlo, J.J., Cox, D.D.: Untangling invariant object recognition. Trends in cognitive sciences **11**(8), 333–341 (2007) 2
11. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021) 2
12. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Dynamic novel-view synthesis: A reality check. In: NeurIPS (2022) 2, 7, 8, 9
13. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives (2024), https://arxiv.org/abs/2311.18259 7, 11

14. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: a scalable dataset generator (2022) 7, 8, 9, 11
15. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: CVPR (2020) 7, 11
16. Guo, Y., Shi, L.X., Chen, J., Finn, C.: Ctrl-world: A controllable generative world model for robot manipulation (2025), https://arxiv.org/abs/2510.10125 2
17. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024), https://arxiv.org/abs/2404.02101 4
18. Hodge, W.V.D., Pedoe, D.: Methods of Algebraic Geometry, Volume 1. Cambridge University Press, London/New York (1947), original Publication 6
19. Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., Ondruska, P.: One thousand and one hours: Self-driving motion prediction dataset. In: CoRL. vol. 155, pp. 409–418 (2020) 7, 11
20. Huang, J., Miao, S., Yang, B., Ma, Y., Liao, Y.: Vivid4d: Improving 4d reconstruction from monocular video by video inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12592–12604 (2025) 2
21. Kahneman, D., Treisman, A., Gibbs, B.J.: The reviewing of object files: Object-specific integration of information. Cognitive psychology **24**(2), 175–219 (1992) 2
22. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Rota Bulò, S., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025), https://arxiv.org/abs/2509.13414 13
23. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., Fagan, P.D., Hejna, J., Itkina, M., Lepert, M., Ma, Y.J., Miller, P.T., Wu, J., Belkhale, S., Dass, S., Ha, H., Jain, A., Lee, A., Lee, Y., Memmel, M., Park, S., Radosavovic, I., Wang, K., Zhan, A., Black, K., Chi, C., Hatch, K.B., Lin, S., Lu, J., Mercat, J., Rehman, A., Sanketi, P.R., Sharma, A., Simpson, C., Vuong, Q., Walke, H.R., Wulfe, B., Xiao, T., Yang, J.H., Yavary, A., Zhao, T.Z., Agia, C., Baijal, R., Castro, M.G., Chen, D., Chen, Q., Chung, T., Drake, J., Foster, E.P., Gao, J., Guizilini, V., Herrera, D.A., Heo, M., Hsu, K., Hu, J., Irshad, M.Z., Jackson, D., Le, C., Li, Y., Lin, K., Lin, R., Ma, Z., Maddukuri, A., Mirchandani, S., Morton, D., Nguyen, T., O'Neill, A., Scalise, R., Seale, D., Son, V., Tian, S., Tran, E., Wang, A.E., Wu, Y., Xie, A., Yang, J., Yin, P., Zhang, Y., Bastani, O., Berseth, G., Bohg, J., Goldberg, K., Gupta, A., Gupta, A., Jayaraman, D., Lim, J.J., Malik, J., Martín-Martín, R., Ramamoorthy, S., Sadigh, D., Song, S., Wu, J., Yip, M.C., Zhu, Y., Kollar, T., Levine, S., Finn, C.: Droid: A large-scale in-the-wild robot manipulation dataset (2024), https://arxiv.org/abs/2403.12945 7, 11
24. Komatsu, H.: The neural mechanisms of perceptual filling-in. Nature reviews neuroscience **7**(3), 220–231 (2006) 2
25. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6165–6177 (2025) 8
26. Liang, J., Liu, R., Ozguroglu, E., Sudhakar, S., Dave, A., Tokmakov, P., Song, S., Vondrick, C.: Dreamitate: Real-world visuomotor policy learning via video generation. arXiv preprint arXiv:2406.16862 (2024), https://arxiv.org/abs/2406.16862 2

27. Lin, K.E., Xiao, L., Liu, F., Yang, G., Ramamoorthi, R.: Deep 3d mask volume for view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1749–1758 (2021) 8
28. Nanay, B.: The importance of amodal completion in everyday perception. i-Perception **9**(4), 2041669518788887 (2018) 2
29. NVIDIA: Cosmos-predict2: Diffusion-based world foundation models for physics-aware image and video generation. https://github.com/nvidia-cosmos/cosmos-predict2 (Jun 2025), accessed: 2025-11-06 8
30. NVIDIA, Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., Chattopadhyay, P., Chen, M., Chen, Y., Chen, Y., Cheng, S., Cui, Y., Diamond, J., Ding, Y., Fan, J., Fan, L., Feng, L., Ferroni, F., Fidler, S., Fu, X., Gao, R., Ge, Y., Gu, J., Gupta, A., Gururani, S., El Hanafi, I., Hassani, A., Hao, Z., Huffman, J., Jang, J., Jannaty, P., Kautz, J., Lam, G., Li, X., Li, Z., Liao, M., Lin, C.H., Lin, T.Y., Lin, Y.C., Ling, H., Liu, M.Y., Liu, X., Lu, Y., Luo, A., Ma, Q., Mao, H., Mo, K., Nah, S., Narang, Y., Panaskar, A., Pavao, L., Pham, T., Ramezanali, M., Reda, F., Reed, S., Ren, X., Shao, H., Shen, Y., Shi, S., Song, S., Stefaniak, B., Sun, S., Tang, S., Tasmeen, S., Tchapmi, L., Tseng, W.C., Varghese, J., Wang, A.Z., Wang, H., Wang, H., Wang, H., Wang, T.C., Wei, F., Xu, J., Yang, D., Yang, X., Ye, H., Ye, S., Zeng, X., Zhang, J., Zhang, Q., Zheng, K., Zhu, A., Zhu, Y.: World simulation with video foundation models for physical ai (2025), https://arxiv.org/abs/2511.00062 3, 6
31. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12999–13008 (2023) 7, 11
32. Pang, J.C., Tang, N., Li, K., Tang, Y., Cai, X.Q., Zhang, Z.Y., Niu, G., Sugiyama, M., Yu, Y.: Learning view-invariant world models for visual robotic manipulation. In: The Thirteenth International Conference on Learning Representations (2025) 2
33. Park, S., Intraub, H., Yi, D.J., Widders, D., Chun, M.M.: Beyond the edges of a view: boundary extension in human scene-selective visual cortex. Neuron **54**(2), 335–342 (2007) 2
34. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., Newell, A., Law, H., Goyal, A., Yang, K., Deng, J.: Infinite photorealistic worlds using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12630–12641 (2023) 7
35. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., Ma, Z., Deng, J.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21783–21794 (June 2024) 7
36. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 4, 6, 9, 10, 11, 14
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3
38. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science **171**(3972), 701–703 (1971) 2

39. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding (2021), https://arxiv.org/abs/2104.09864 6
40. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020) 7, 11
41. Tian, S., Wulfe, B., Sargent, K., Liu, K., Zakharov, S., Guizilini, V., Wu, J.: View-invariant policy learning via zero-shot novel view synthesis. arXiv preprint arXiv:2409.03685 (2024), https://arxiv.org/abs/2409.03685 2
42. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019), https://arxiv.org/abs/1812.01717 12
43. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. ECCV (2024) 4, 6, 7, 8, 9, 10, 11, 12, 14
44. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision. pp. 439–457. Springer (2024) 4
45. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025), https://arxiv.org/abs/2503.20314 3
46. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: International Conference on Computer Vision (ICCV) (2025) 2, 4, 8, 9
47. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024) 2
48. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021) (2021) 7, 11
49. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024) 2
50. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26057–26068 (2025) 4
51. Xiao, Z., Ouyang, W., Zhou, Y., Yang, S., Yang, L., Si, J., Pan, X.: Trajectory attention for fine-grained video motion control. In: ICLR (2025) 4, 6, 9, 10, 11, 14
52. Yan, Y., Xu, Z., Lin, H., Jin, H., Guo, H., Wang, Y., Zhan, K., Lang, X., Bao, H., Zhou, X., Peng, S.: Streetcrafter: Street view synthesis with controllable video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025) 4
53. Yang, J., Huang, I., Vu, B., Bajracharya, M., Antonova, R., Bohg, J.: Mobi-pi: Mobilizing your robot learning policy. arXiv preprint arXiv:2505.23692 (2025), https://arxiv.org/abs/2505.23692 2

54. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv:2406.09414 (2024), https://arxiv.org/abs/2406.09414 12, 14
55. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024), https://arxiv.org/abs/2408.06072 3
56. Yesiltepe, H., Yanardag, P.: Dynamic view synthesis as an inverse problem (2025), https://arxiv.org/abs/2506.08004 3, 8
57. Yesiltepe, H., Yanardag, P.: Dynamic view synthesis as an inverse problem. In: NeurIPS (2025) 4
58. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5336–5345 (2020) 8
59. YU, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In: Int. Conf. Comput. Vis. (2025) 3, 4, 6, 9, 10, 11, 14
60. Zhang, D.J., Paiss, R., Zada, S., Karnad, N., Jacobs, D.E., Pritch, Y., Mosseri, I., Shou, M.Z., Wadhwa, N., Ruiz, N.: Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. arXiv preprint arXiv:2411.05003 (2024), https://arxiv.org/abs/2411.05003 8, 9
61. Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., Li, X.: Cami2v: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv:2410.15957 (2024), https://arxiv.org/abs/2410.15957 4
62. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In: ICCV (2023) 7