Generative Camera Dolly: Extreme Monocular Dynamic Novel View Synthesis

Basile Van Hoorick[1], Rundi Wu[1], Ege Ozguroglu[1], Kyle Sargent[2], Ruoshi Liu[1], Pavel Tokmakov[3], Achal Dave[3], Changxi Zheng[1], and Carl Vondrick[1]

[1] Columbia University
[2] Stanford University
[3] Toyota Research Institute

gcd.cs.columbia.edu

Abstract. Accurate reconstruction of complex dynamic scenes from just a single viewpoint continues to be a challenging task in computer vision. Current dynamic novel view synthesis methods typically require videos from many different camera viewpoints, necessitating careful recording setups, and significantly restricting their utility in the wild as well as in terms of embodied AI applications. In this paper, we propose **GCD**, a controllable monocular dynamic view synthesis pipeline that leverages large-scale diffusion priors to, given a video of any scene, generate a synchronous video from any other chosen perspective, conditioned on a set of relative camera pose parameters. Our model does not require depth as input, and does not explicitly model 3D scene geometry, instead performing end-to-end video-to-video translation in order to achieve its goal efficiently. Despite being trained on synthetic multi-view video data only, zero-shot real-world generalization experiments show promising results in multiple domains, including robotics, object permanence, and driving environments. We believe our framework can potentially unlock powerful applications in rich dynamic scene understanding, perception for robotics, and interactive 3D video viewing experiences for virtual reality.

1 Introduction

Video generation has made tremendous progress in recent years. Results from Sora [7], OpenAI's recently released text-to-video generation model, have shown that generating a high-quality video as long as one minute is possible. Following the scaling curve, video models will most likely continue to improve in many aspects. However, one essential capability is still missing from these video models to be useful for many downstream applications – the ability to generate the same dynamic scene from an arbitrary camera perspective based on an existing video.

In this paper, we aim to tackle the problem of *dynamic novel view synthesis* (DVS) – given a video of a dynamic scene, we aim to generate a video of the same scene from another viewpoint. Once we develop a solution for this problem, we

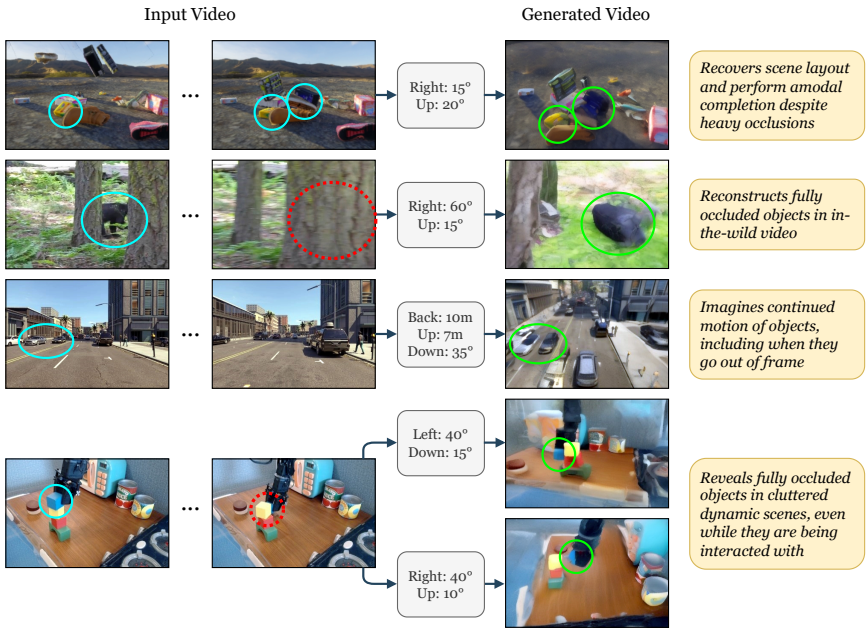

Fig. 1: Spatial video translation of dynamic scenes. Given a single RGB video, we propose a method that is capable of imagining what that scene would look like from another viewpoint. Even for extreme camera transformations with large angles, our approach synthesizes videos with rich visual details that are consistent with the input, demonstrating advanced spatiotemporal reasoning capabilities.

can leverage it for several impactful use cases, such as generating novel views of a live street scenario based on cameras mounted on an autonomous vehicle; seeing a cluttered environment from a different viewpoint while a robot is performing dexterous manipulations; enabling geometrically consistent video passthrough for mixed reality [78]; and immersively reliving videos recorded in the past.

However, this task is naturally extremely ill-posed and challenging. While yielding promising results, prior works typically addressed it by assuming either that contemporaneous multi-viewpoint video is available [38,47,71,75,80], and/or by imposing that the relative camera viewpoint changes must be small (*i.e.* limited to just a handful of degrees) [32,69]. These restrictions make them vastly insufficient for the aforementioned applications, which require in-the-wild novel view synthesis pipelines with dramatic camera viewpoint changes.

Free-viewpoint synthesis from a single video requires prior knowledge because it is highly under-constrained. Modern video generative models, such as Stable Video Diffusion [5], have learned rich priors for real-world dynamics, 3D geometry, and camera motions, as they are trained on hundreds of millions of video clips from the Internet. In this work, we propose an approach to capitalize on these rich representations for the task of DVS. We curate pairs of videos

of dynamic scenes from simulation as training data, and apply them to steer a pretrained video generative model towards the desired behavior by finetuning.

Qualitative and quantitative results demonstrate that our model achieves state-of-the-art results on the task of monocular DVS, and generalizes effectively to various out-of-distribution scenes, including real-world driving videos, robot manipulation scenes, and other in-the-wild videos with heavy occlusion patterns, as shown in Figure 1. Much like a camera dolly in film-making [74], our approach essentially conceives a virtual camera that can move around with up to six degrees of freedom, reveal significant portions of the scene that are otherwise unseen, reconstruct hidden objects behind occlusions, all within complex dynamic scenes, even when the contents are moving.

Our core contribution is the design and evaluation of a framework, *Generative Camera Dolly* (GCD), for learning to generate videos from novel viewpoints of a dynamic scene, using an end-to-end video-to-video neural network. Section 2 provides a brief overview of related work. Section 3 introduces the approach including the model architecture, and a description of how to achieve precise camera control within the video diffusion model. Section 4 discusses training data, benchmarks, and task details. Section 5 investigates important hyperparameter decisions with regard to the conceptual implementation of camera control. Section 6 provides both quantitative and qualitative evaluation of the system as well as several examples of our model generalizing to out-of-distribution data. We believe the ability to perform free-viewpoint video synthesis for a dynamic scene will have a significant impact on 3D/4D computer vision research, as well as other related areas, including content creation, AR/VR, and robotics.

2 Related Work

Dynamic scene reconstruction. The landscape of dynamic scene novel view synthesis has been primarily dominated by techniques that rely on multiple synchronized (*i.e.* contemporaneous) input videos [2, 4, 8, 30, 47, 71, 80, 85], which limits their practical usage in real-world scenarios. The emergence of Neural Radiance Fields (NeRF) [41] has catalyzed a revolution in dynamic view synthesis, presenting state-of-the-art results in this domain [14, 31, 43, 44, 48, 62, 77]. Most such methods represent scenes through time-evolving NeRFs [10, 16, 17, 31, 68, 77], for handling complicated, dynamic 3D scene motions in casual videos.

A notable trend in recent advancements involves the synthesis of novel views from a single camera perspective [17, 32, 36, 69, 81, 87]. DynIBaR adopts a volumetric image-based rendering framework that, instead of encoding and compressing the entire scene within a single representation (for example an MLP), aggregates features from nearby views in a camera motion-aware manner, which enables synthesizing novel views for long videos with uncontrolled camera paths [32]. DpDy leverages an image-based diffusion model to iteratively distill knowledge coming from diffusion priors into a hybrid 4D representation [69].

It is worth noting that essentially all aforementioned methods optimize *per-scene* representations independently of each other. Therefore, they are (1) largely

unable to share any knowledge between different reconstructions, such as to generalize to unseen environments; and (2) largely unable to infer or extrapolate from incomplete observations, such as to recover fully occluded regions. Moreover, failure modes are often observed when the monocular input video lacks *effective* multi-view cues, for example as enabled implicitly thanks to a moving, especially a fast-moving, camera [17].

Video diffusion models. Recent work has rapidly improved the state of video generation models. Most generative models focus on diffusion-based approaches [5, 6,18,24,26,57], though important exceptions exist, particularly with autoregressive training [73, 83]. Following recent work which shows image-based diffusion models can be re-purposed for computer vision tasks including monocular depth estimation [53], 3D reconstruction [35] and amodal segmentation [42], our work adopts a public video diffusion model for dynamic view synthesis. We rely on Stable Video Diffusion [5] as it generates high quality videos, and provides a public image-to-video model checkpoint with code, although our framework can generalize to any video generation approach.

3D and 4D generation. Most of the works enabling successful 3D generation via generative models hence rely on channelling the representational power of 2D diffusion models towards a single 3D representation that is iteratively optimized over time, for example through score distillation [46]. This *multiview 2D-to-3D* paradigm is exemplified by many text-to-3D and image-to-3D works [11,23,27,33, 35,45,46,66,70,72,76,86]. Emphasizing the temporal component, text-to-4D and image-to-4D papers have begun appearing as well, although the results currently remain mostly limited to animations of single objects or animals [1, 34, 58, 88]. Video-to-4D, which is likely harder because every frame of the observation must be respected, has remained less explored so far. In [64], a video-to-4D scene reconstruction task and framework is proposed, although the model requires depth input, and only works in narrow domains as it is trained from scratch.

Object permanence and amodal completion. The problem of reasoning about the invisible parts of a scene has been studied extensively in the literature, but so far almost exclusively from an object-centric perspective. For example, in the image world, amodal completion [15,42,84] studies the problem of reconstructing the occluded parts of an object based on its visible parts and the scene context. However, these methods are naturally restricted to partial occlusions. In contrast, for videos, some object tracking methods capitalize on the temporal context to reason about the location [55,60,61] or even shape [65] of fully occluded instances.

While abstracting the full complexity of a dynamic scene into a compact set of objects allows these methods to be relatively data- and compute-efficient, it also limits their applicability. In this work, we propose a more general approach that is capable of revealing any parts of a scene, together with their dynamics, similar to [64]. This includes not only occluded objects, but also 'stuff' regions [9], such as natural or man-made surfaces, liquids, and so on.

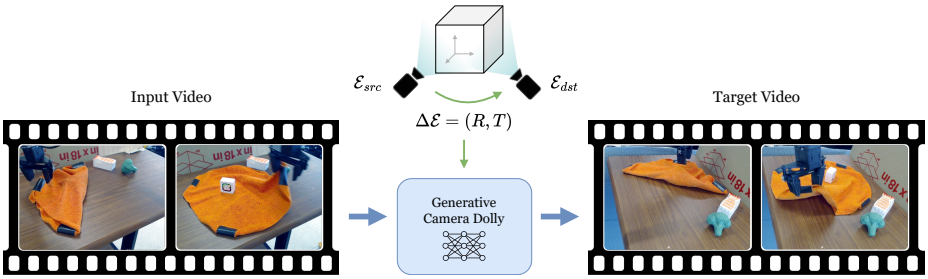

Fig. 2: Method. Our model, GCD, is an end-to-end video translation pipeline that maps an input video from any viewpoint into an output video from any other perspective, with the objective of respecting all objects and dynamics occurring within the observed dynamic scene, and faithfully reconstructing the corresponding visual details from this novel viewpoint. The relative camera extrinsics matrix $\Delta\mathcal{E}$ guides the relationship between the two camera poses.

We note that at least one concurrent work also tackles dynamic view synthesis: in Exo2Ego [39], authors propose a framework that translates third-person (exocentric) to first-person (egocentric) videos on a per-frame basis, incorporating priors for hand-object interactions and focusing primarily on those scenarios.

3 Approach

First, we formally introduce the task of monocular dynamic novel view synthesis from unconstrained video input. Let $\boldsymbol{x} \in \mathbb{R}^{T\times H\times W\times 3}$ be RGB frames captured from a single camera perspective, that encode the visual observation of a dynamic scene of interest. We denote its associated camera extrinsics matrix as $\mathcal{E}_{src} \in \mathbb{R}^{T\times 4\times 4}$, and define $\mathcal{E}_{dst} \in \mathbb{R}^{T\times 4\times 4}$ to be the desired target camera extrinsics matrix over time. Our model f is then tasked with predicting a video $\boldsymbol{y} \in \mathbb{R}^{T\times H\times W\times 3}$, that plausibly depicts the same dynamic scene from the specified new viewpoint. For simplicity, and without loss of generality, we assume that (1) the output video is temporally synchronized with the input, and (2) the camera intrinsics matrix $\mathcal{K} \in \mathcal{R}^{3\times 3}$ stays constant over time as well as across pose changes; notably, $\boldsymbol{y}$ assumes the same focal length as $\boldsymbol{x}$.

Since novel view synthesis is an inherently under-constrained, challenging problem, our approach will use existing large-scale video generative models. Diffusion models have been shown to excel at image-to-3D tasks [35, 37, 56, 76], justifying our attempt to perform video-to-4D. Moreover, they have shown remarkable zero-shot abilities in generating realistic, diverse videos from user-given text descriptions and/or initial frames [3, 5, 26]. However, they are typically not trained to accept video as a conditioning signal, and fine-grained control over camera transformations is also not available by default. To overcome these obstacles, we must make a few architectural changes.

3.1 Camera viewpoint control

Given a single RGB video $\boldsymbol{x}$ of a dynamic scene, our goal is to synthesize another video $\boldsymbol{y}$ of the scene from a different viewpoint. Since large-scale video diffusion models have been trained on hyper-scale data, their support of the natural video distribution most likely covers a wide range of realistic scenes and viewpoints. To this end, given a dataset of paired videos and their *relative* camera extrinsics $\Delta\mathcal{E} = \{\mathcal{E}_{src,t}^{-1} \cdot \mathcal{E}_{dst,t}\}_{t=0}^{T-1} \in \mathbb{R}^{T\times 4\times 4}$ over time, we teach a latent diffusion model f to learn controls over camera parameters within any video $\boldsymbol{x}$:

$$\boldsymbol{y} = f\left(\boldsymbol{x}, \Delta\mathcal{E}\right) \tag{1}$$

Specifically, we modify Stable Video Diffusion (SVD) to accept a new form of *micro-conditioning*, a term coined in [5], which is designed for the purpose of communicating low-dimensional metadata (such as the desired frame rate of the output video, and the amount of optical flow) to the network. We decompose $\Delta\mathcal{E}_t \in \mathrm{SE}(3)$ into a series of camera rotation matrices $R_t \in \mathrm{SO}(3)$ and translation matrices $T_t \in \mathbb{R}^3$ over time, project this flattened information through an MLP m, and add the resulting embedding to the feature vectors at various convolutional layers placed throughout the network, similarly to the concurrent work SV3D [66]. The diffusion timestep, FPS, and motion strength are also passed to the network this way. To preserve the existing priors of SVD as much as possible, we initialize the network weights based on the publicly available image-to-video model checkpoint. The new embedder m that processes $\{(R_t, T_t)\}$ is randomly initialized with default parameters. After training the network end-to-end, the resulting model is capable of imagining unseen videos from any chosen perspective, as illustrated in Figure 2 (high-level) and Figure **??** (detail).

3.2 Video conditioning

To accurately perform dynamic view synthesis, both low-level perception (to analyze the visible geometry, shapes, appearance, etc.) and high-level understanding (to infer the occluded regions, based on world knowledge as well as other observed frames) of the input video is required. We adopt the same hybrid conditioning mechanism as SVD [5], where the visual signal is processed in two ways. In case of image-to-video, the first stream calculates the CLIP [49] embedding $c(\boldsymbol{x}_0)$ of the incoming image to condition the U-Net ϵ via cross-attention, and the second stream channel-concatenates the VAE-encoded image $\boldsymbol{x}_0$ with all frames of the video sample $\hat{\boldsymbol{y}}$ that are being denoised.

We keep this mechanism almost entirely intact when moving from the pretraining to the finetuning stage, but we propose to simply substitute the first frame $\boldsymbol{x}_0$ for the entire input video $\boldsymbol{x}$ from the source viewpoint, such that the conditioning information now becomes a function of time. This ensures that our model has the opportunity to watch how the dynamic scene unfolds over time, and hence must learn to respect the dynamics and physics of the objects within.

In architectural terms, the output sample $\hat{\boldsymbol{y}}$ has contemporaneous input frames from $\boldsymbol{x}$ attached to it for every video timestamp t, such that at diffusion

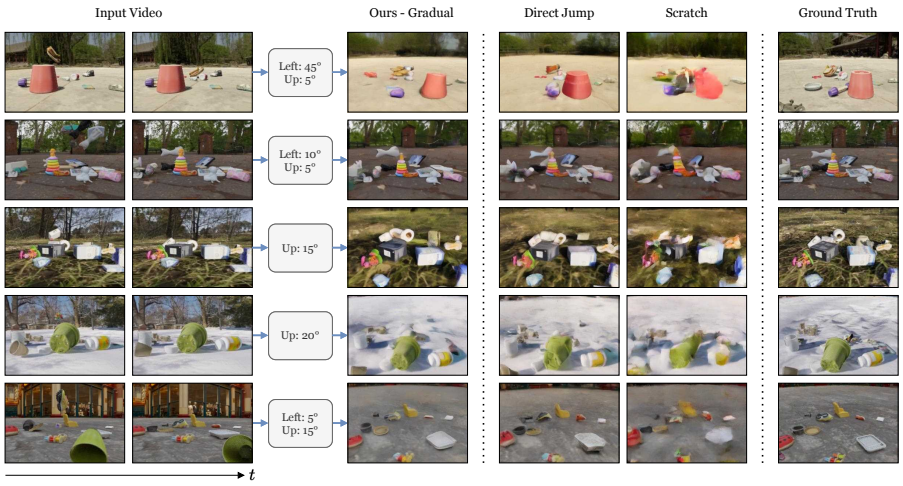

Fig. 3: Qualitative ablation study results for Kubric-4D. We show inputs, predictions, ablations, and ground truths. The input and output videos both consist of $T = 14$ frames, but we show the first and last frame of the input video for conciseness, and only the last frame of the output and target. Whereas the ablations tend to look blurry with incorrect shape and/or appearance characteristics (especially for moving objects), our main model (gradual, max 90°, finetuned) faithfully reconstructs the scene layout and dynamics from the input video. In addition, it often hallucinates plausible backgrounds in unseen regions.

noise timestep u during inference, where $w \in [1, \infty)$ is the guidance strength for classifier-free guidance [25]:

$$\hat{\boldsymbol{y}}_{u-1} = w\epsilon\left(\hat{\boldsymbol{y}}_u \parallel \boldsymbol{x}, \Delta\mathcal{E}\right) - (w-1)\epsilon\left(\hat{\boldsymbol{y}}_u\right), \tag{2}$$

4 Datasets

While the availability of multi-view video data has been growing [13, 19, 20, 29, 50, 51, 54, 63, 64, 90], it is still relatively sparse compared to conventional image or video datasets. In order to train and evaluate our model, we require a decent amount of multi-view RGB videos from highly cluttered dynamic scenes. To this end, we contribute two high-quality synthetic datasets, shown in Figures 3, 4, and 5 and briefly describe them below.

4.1 Kubric-4D

We leverage the Kubric [21] simulator as our data source for generic multi-object interaction videos, carrying a high degree of visual detail and physical realism. Each scene contains between 7 and 22 randomly sized objects in total, with roughly one-third of them spawned in mid-air at the beginning of the video

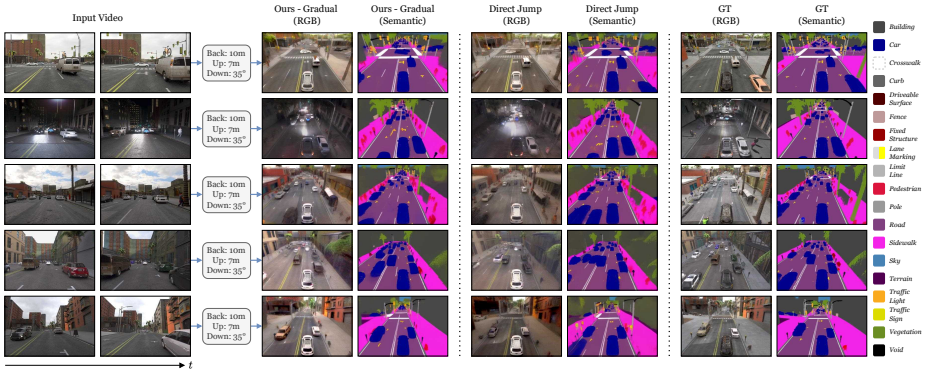

Fig. 4: Qualitative ablation study results for ParallelDomain-4D. We show inputs, predictions, ablations, and ground truths for both visual and semantic scene completion. Our model excels at recovering the top-down viewpoint with high accuracy in both modalities, despite the heavy occlusion patterns that often occur in driving scenes. While the *direct* model performs almost as well as the *gradual* one, it tends to introduce slightly more hallucination and discoloration of objects.

to encourage sophisticated dynamics. Complicated occlusion patterns arise very frequently, making this dataset highly challenging for accurate novel view synthesis. We generate 3,000 scenes of 60 frames each, at a frame rate of 24 FPS, with RGB-D data rendered from 16 virtual cameras at a fixed set of poses.

Because the dynamic scene is sufficiently densely covered, we unproject all the pixels from available viewpoints into a merged 3D point cloud per frame. As a form of data augmentation, we then render them into videos from arbitrary viewpoints according to camera trajectories that can be chosen and controllably sampled depending on the exact training configuration.

4.2 ParallelDomain-4D

Since rich scene understanding and spatial reasoning skills are paramount for maximizing situational awareness in the context of driving, we employ the state-of-the-art data generation service ParallelDomain to produce complex, highly photorealistic road scenes. The videos depict driving scenarios covering a wide variety of locations, vehicles, persons, traffic situations, and weather conditions. Here, we have 1,533 scenes available of 50 frames each, at a frame rate of 10 FPS, with high-quality annotations for multiple modalities (RGB colors, semantic categories, instance IDs, etc.) along with per-pixel ground truth depth rendered from 19 virtual cameras at a fixed set of poses.

In our experiments, we train separate models for RGB view synthesis and semantic view synthesis; the latter demonstrates that the predicted modality need not be the same as the given modality. Similarly as for Kubric-4D, we perform a unproject-and-reproject routine to turn this multi-view video dataset into a pseudo-4D data source from which we can render videos of the scene from arbitrary camera perspectives, within pre-defined spatiotemporal bounds.

Variant	PSNR (all) ↑	SSIM (all) ↑	LPIPS (all) ↓	PSNR (occ.) ↑	SSIM (occ.) ↑
Ours (direct, max 90°, scratch)	15.96	0.450	0.575	15.85	0.470
Ours (direct, max 180, scratch)	14.71	0.426	0.611	15.15	0.458
Ours (gradual, max 90°, scratch)	16.92	0.486	0.542	16.59	0.494
Ours (gradual, max 180°, scratch)	16.63	0.479	0.552	16.34	0.491
Ours (direct, max 90°, finetuned)	17.23	0.494	0.507	16.69	0.492
Ours (direct, max 180°, finetuned)	16.65	0.471	0.529	16.18	0.470
Ours (gradual, max 90°, finetuned)	**17.88**	**0.521**	**0.486**	**17.33**	0.514
Ours (gradual, max 180°, finetuned)	17.81	**0.521**	0.488	17.20	**0.515**

Table 1: Ablation study results on Kubric. We evaluate various versions of our dynamic view synthesis model on only the last frame for fairness, *i.e.* to ensure that the direct and gradual trajectory models are spatially aligned. See Figure 3 for qualitative illustrations.

4.3 Task details

In our experiments, without loss of generality, we assume a static input camera pose $\mathcal{P}_{src}$,[4] and pick a target destination pose $\mathcal{P}_{dst}$ that we want the output camera to reach at or before the end of the generated video. In general, $\mathcal{P}$ is a *pose description* that can be defined in many ways, for example a set of spherical coordinates that represent the camera position and look-at location, but (1) must allow for convex interpolation (*i.e.* $\alpha\mathcal{P}_1 + (1-\alpha)\mathcal{P}_2$ with $\alpha \in [0, 1]$ is valid), and (2) is connected to a valid 6-DoF rigid body transformation $\mathcal{E} \in \mathrm{SE}(3)$ through the function g, *i.e.* $\mathcal{E} = g(\mathcal{P})$.

When training for the task of dynamic view synthesis on Kubric-4D, pairs of input and output poses are randomly sampled within certain spherical coordinate bounds (both in absolute terms and relative to each other), with the extra condition that they are looking at the center of the 3D scene, *i.e.* $\mathcal{P} \in \mathbb{R}^3$.

In case of ParallelDomain-4D, the input video and pose always correspond to the ego vehicle's forward-facing viewpoint, as if a sensor were mounted on the front of the car. The output pose is a fixed top-down viewpoint with the ego vehicle at the bottom center, which enables a detailed overview of surroundings.

5 Choice of camera trajectory

Our formulation of the dynamic view synthesis task in Section 3 is quite general, so it is worth thinking about which specific instantiations of this conceptual framework would be most effective in practice. Given arbitrary video inputs, our goal is to devise a structured protocol for choosing relative camera trajectories that both maximize the exploitation of knowledge contained within the pre-trained SVD representation, as well as enable a detailed understanding of the

[4] This can always be achieved by defining the reference coordinate system to move along with the recording camera.

Variant	PSNR (all) ↑	SSIM (all) ↑	LPIPS (all) ↓	PSNR (occ.) ↑	SSIM (occ.) ↑
Ours (direct, scratch)	22.49	0.622	0.487	22.62	0.653
Ours (gradual, scratch)	22.73	0.632	0.467	22.76	0.664
Ours (direct, finetuned)	23.32	0.664	0.440	23.29	0.691
Ours (gradual, finetuned)	**23.47**	**0.670**	**0.425**	**23.52**	**0.696**

Table 2: Ablation study results on ParallelDomain in RGB space. We perform visual scene completion, and evaluate various dynamic view synthesis models on only the last frame for fairness, similarly to Table 1. See Figure 4 for qualitative illustrations.

Variant	mIoU (all) ↑	mIoU (occ.) ↑
Ours (direct, from scratch)	31.2%	28.6%
Ours (gradual, from scratch)	34.4%	32.1%
Ours (direct, finetuned)	36.7%	35.4%
Ours (gradual, finetuned)	**39.0%**	**37.7%**

Table 3: Ablation study results on ParallelDomain in semantic space. We perform semantic completion of the scene, again similarly to Table 1. See Figure 4 for qualitative illustrations.

dynamic scene observed at inference time to the fullest extent possible. Specifically, we wish to synthesize views that reach as far as the opposite end of the scene, *e.g.* by orbiting the azimuth angle ϕ up to 180°. This is considerably more dramatic than what the state of the art in dynamic view synthesis is typically capable of [10,17,32,75], and allows us to reveal large, formerly unseen portions.

However, it turns out that opposing forces are at play. On one hand, we wish to get to the destination camera pose *"as fast as possible"* (because the scene could already be evolving and changing over time as we are watching it). On the other hand, if the output video moves away from the source viewpoint too quickly, we might risk incurring a *distribution misalignment* due to the fact that the image-to-video SVD model predominantly generates videos that start at nearly the exact same spatial perspective as the given image. Moreover, the camera generally does not move much throughout the video, typically performing only minor panning motions and/or mild rotations.

To resolve this concern, we translate it into three questions: (1) where should the output pose *start*; (2) how fast should it be taught to *move* in-between subsequent frames; and (3) how much does finetuning, *i.e.* borrowing priors from SVD help (or hurt) in each case, versus training an identical network from scratch? We investigate this by running comparative studies on both the Kubric-4D and ParallelDomain-4D datasets. For each tested scene, we fix a source pose $\mathcal{P}_{src}$ and a destination pose $\mathcal{P}_{dst}$, following Section 4.3. Using $\mathcal{E}_{src,t} = g(\mathcal{P}_{src})$ and $\alpha = \frac{t}{T-1} \in [0, 1]$, we define *gradual* and *direct* trajectories as follows:

$$\mathcal{E}_{dst,t} = \begin{cases} g\left(\alpha \mathcal{P}_{dst} + (1-\alpha)\mathcal{P}_{src}\right), & \forall t, \quad \text{if gradual} \\ g\left(\mathcal{P}_{dst}\right), & \forall t, \quad \text{if direct} \end{cases} \tag{3}$$

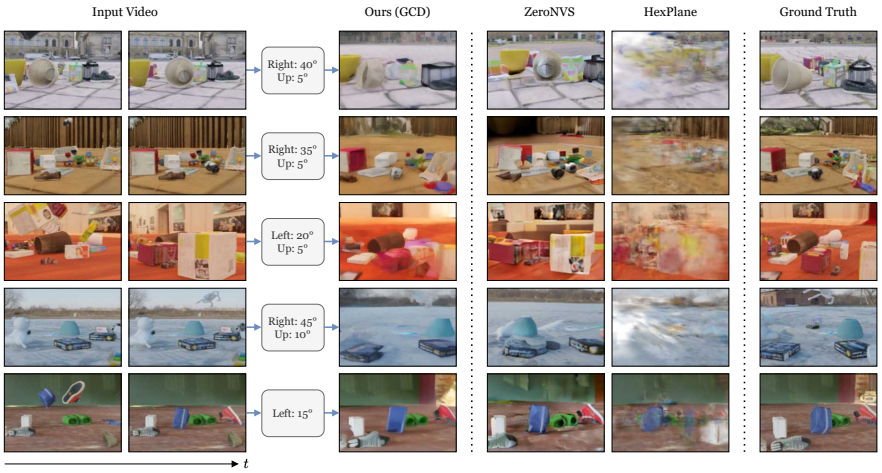

Fig. 5: Qualitative baseline comparison results for Kubric-4D. We show inputs, predictions, baselines, and ground truths. Compared to baselines, our results depict the scene layout and dynamics under the desired novel viewpoints with reasonable accuracy overall and much fewer flickering artefacts.

In other words, *gradual* means that the virtual camera pose corresponding to the output video linearly interpolates (in an intermediate description space, for example spherical coordinates) between $\mathcal{P}_{src}$ and $\mathcal{P}_{dst}$ from start to end, whereas *direct* implies that the generated video directly adheres precisely to $\mathcal{P}_{dst}$ at every frame without interpolation. For Kubric-4D, *max 90°* limits the relative horizontal (*i.e.* azimuth) angle variation between input and output to $\pm 90°$ at training time ($|\Delta\phi| \leq 90°$), whereas *max 180°* effectively allows for synthesizing any 360°-surround viewpoint of the dynamic scene.

The results are shown in Tables 1, 2, and 3, and Figures 3 and 4. From this ablation study, we observe with Kubric-4D that: (1) it is preferable to gradually interpolate from source to destination pose than to immediately jump there (+1.17 dB average PSNR improvement between *direct* and *gradual*); (2) there exists a trade-off between the range of camera transformations the model should be trained for, and how extreme of a rotation one wishes to be able to achieve at most (+0.55 dB between *max 180°* and *max 90°*); and (3) it is preferable to start from the SVD checkpoint that had been trained on large-scale video rather than to train from random initialization, though not by a particularly huge margin (+1.34 dB between *scratch* and *finetuned*). Although *gradual, max 90°, finetuned* and *gradual, max 180°, finetuned* are very close, we proceed with the former in all further experiments, described below. We make consistent findings in the ParallelDomain-4D dataset, where *gradual, finetuned* is the best model.

6 Experiments

In this section, we evaluate our monocular dynamic novel view synthesis framework. We adopt the SVD variant that predicts $T = 14$ frames, but due to

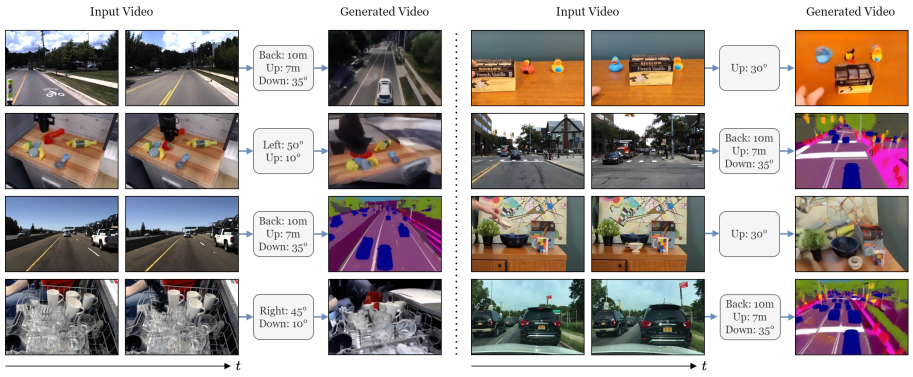

Fig. 6: Qualitative real-world generalization results. We show inputs and predictions on BridgeData V2 [67], TCOW Rubric [65], TRI-DDAD [22], and Berkeley DeepDrive [82]. Despite being trained on synthetic data alone, our approach show surprisingly strong generalization skills to a variety of real-world scenarios. For example, on the top right, where a full occlusion occurs around the middle of the video, our model faithfully predicts both the position and appearance of the invisible duck at the last frame, demonstrating object permanence capabilities.

computational constraints, we downscale the input and output resolution to $W \times H = 384 \times 256$. We report numerical results on the test splits of our two in-domain datasets (Kubric-4D and ParallelDomain-4D), comparing against several state-of-the-art baselines, but additionally showcase promising qualitative results on in-the-wild videos from various domains. For more results as well as animated visualizations, please see gcd.cs.columbia.edu.

6.1 Evaluation metrics

Following related work in novel view synthesis [28, 32, 35, 40, 52, 69], for predictions in RGB space, we evaluate PSNR, SSIM, and LPIPS scores and average the results across video frames and test examples. For semantic category predictions, following conventions in semantic segmentation [12, 59, 79, 89], we first calculate the average Intersection over Union (IoU) per category over the whole ParallelDomain test set, and then report the mean IoU (mIoU) across categories.

Based on the ground truth input viewpoint depth map, it is also possible to determine which pixels in the target viewpoint are visible or hidden. In addition to the regular metrics ("all"), we therefore spatially mask the videos to determine metrics for occluded regions only ("occ."), which the model has to inpaint.

Even though our model accepts and predicts the same number of frames ($T = 14$), the first output frame for the *gradual* camera trajectory models (described below) is spatially aligned with the first input frame. This implies that it could in principle be solved by copying its pixels (except if the task involves switching to another modality, for example semantic category prediction), so we exclude the first frame from the evaluation to avoid inflating the metrics, instead averaging only over the last $T - 1 = 13$ frames, which correspond to different extrinsics.

Method	PSNR (all) ↑	SSIM (all) ↑	LPIPS (all) ↓	PSNR (occ.) ↑	SSIM (occ.) ↑
HexPlane [10]	15.38	0.428	0.568	14.71	0.428
4D-GS [75]	14.92	0.388	0.584	14.55	0.392
DynIBaR [32]	12.86	0.356	0.646	12.78	0.358
Vanilla SVD [5]	13.85	0.312	0.556	13.66	0.326
ZeroNVS [52]	15.68	0.396	0.508	14.18	0.368
Ours	**20.30**	**0.587**	**0.408**	**18.60**	**0.527**
Reproject RGB-D*	12.51	0.537	0.416	-	-

Table 4: Baseline comparison results on Kubric-4D. We evaluate gradual dynamic view synthesis models on all 13 output frames, and with a single RGB video as input. We significantly outperform all baselines for both visible and occluded pixels. *Uses privileged information, *i.e.* can access the ground truth depth map from the input viewpoint.

Method	PSNR (all) ↑	SSIM (all) ↑	LPIPS (all) ↓	PSNR (occ.) ↑	SSIM (occ.) ↑
Vanilla SVD [5]	12.88	0.400	0.658	13.96	0.466
ZeroNVS [52]	18.88	0.490	0.555	19.29	0.552
Ours	**25.04**	**0.731**	**0.358**	**24.70**	**0.733**
Reproject RGB-D*	17.66	0.459	0.441	-	-

Table 5: Baseline comparison results on ParallelDomain in RGB space. We perform visual scene completion, and evaluate gradual dynamic view synthesis on all 13 output frames, and with a single RGB video as input. We significantly outperform all baselines for both visible and occluded pixels. *Uses privileged information, *i.e.* can access the ground truth depth map from the input viewpoint.

6.2 Baselines

We compare our final models against the state-of-the-art dynamic view synthesis methods including *HexPlane* [10], *4D-GS* [75] and *DynIBaR* [32], which all perform per-scene optimization. While these baselines are capable of handling videos with higher resolutions than ours, they are typically limited to much smaller camera angle changes in the one- or low-number-of-views regime, and inference runtimes are many orders of magnitude larger (*e.g.* hours vs. seconds).

In addition, we compare to two pretrained diffusion models *Vanilla SVD* [5] and *ZeroNVS* [52] by adapting them for our task. For Vanilla SVD, we run the original SVD model to generate videos based on the first input frame, without any changes or finetuning. For ZeroNVS, which can generate novel views of scenes based on a single image, we run it for all the input frames independently.

Finally, we compare to a simple geometric baseline (*Reproject RGB-D* and *Reproject Sem-D*), where we reproject pixels from input frames to target viewpoints using the ground truth depth maps, switching to the appropriate modality

Method	mIoU (all) ↑	mIoU (occ.) ↑
Ours	**43.4**%	**38.2**%
Reproject Sem-D*	37.3%	-

Table 6: Baseline comparison results on ParallelDomain in semantic space. We perform semantic completion of the scene, still based on a single RGB video as input. *Uses privileged information, *i.e.* can access the ground truth depth map *and* ground truth semantic category of all input pixels.

as needed. Here, the goal is to study how much information is contained within the input video itself, if precise per-pixel depth values were fully known.

All methods observe the same monocular input video, and are evaluated on the exact same set of randomly sampled output camera trajectories for fairness.

6.3 Results

We report quantitative results in Tables 4, 5, and 6, and show qualitative results in Figures 5 and 4. On both datasets, our model outperforms baseline methods by a large margin. Per-scene optimization methods (e.g., HexPlane) fail to reconstruct the 4D scene representation from a single input view, and thus the rendered videos from novel viewpoints have severe artifacts. Vanilla SVD is able to generate smooth videos but fails to follow the desired camera trajectories, and does not incorporate content from later frames. ZeroNVS can synthesize plausible individual frames from specified viewpoints, but the resulting videos are not temporally coherent and do not respect the scene dynamics. In contrast, our model mostly generates plausible videos that accurately depict the complex scene geometry and motion under the desired novel viewpoint transformations. Apart from the evaluation on in-domain datasets, we also showcase promising results on real-world in-the-wild videos. As shown in Figure 6, our model sometimes generalizes quite well to various domains including driving environments, daily indoor videos, and robotic manipulation scenes.

7 Discussion

In this paper, we present a framework for dynamic novel view synthesis from a monocular video by finetuning a large-scale pretrained video diffusion model [5] on high-quality synthetic data. While we show promising results on real-world in-the-wild videos, our model still struggles on significantly out-of-distribution examples, *e.g.* videos with moving humans. Nevertheless, we believe this work delivers meaningful progress in terms of gaining a rich, detailed understanding of 4D scenes, and takes a solid first step towards enabling zero-shot dynamic view synthesis from a monocular video.

Acknowledgements: This research is based on work partially supported by the NSF CAREER Award #2046910 and the NSF Center for Smart Streetscapes (CS3)

under NSF Cooperative Agreement No. EEC-2133516. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

1. Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L., Wonka, P., Tulyakov, S., Park, J.J., Tagliasacchi, A., Lindell, D.B.: 4d-fy: Text-to-4d generation using hybrid score distillation sampling. arXiv preprint arXiv:2311.17984 (2023)
2. Bansal, A., Vo, M., Sheikh, Y., Ramanan, D., Narasimhan, S.: 4d visualization of dynamic events from unconstrained multi-view videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5366–5375 (2020)
3. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., et al.: Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945 (2024)
4. Bemana, M., Myszkowski, K., Seidel, H.P., Ritschel, T.: X-fields: Implicit neural view-, light-and time-image interpolation. ACM Transactions on Graphics (TOG) **39**(6), 1–15 (2020)
5. 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)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
7. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`
8. Broxton, M., Flynn, J., Overbeck, R., Erickson, D., Hedman, P., Duvall, M., Dourgarian, J., Busch, J., Whalen, M., Debevec, P.: Immersive light field video with a layered mesh representation. ACM Transactions on Graphics (TOG) **39**(4), 86–1 (2020)
9. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018)
10. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
11. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
13. Corona, K., Osterdahl, K., Collins, R., Hoogs, A.: Meva: A large-scale multiview, multimodal video dataset for activity detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1060–1068 (2021)

14. Du, Y., Zhang, Y., Yu, H.X., Tenenbaum, J.B., Wu, J.: Neural radiance flow for 4d view synthesis and video processing. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14304–14314. IEEE Computer Society (2021)
15. Ehsani, K., Mottaghi, R., Farhadi, A.: Segan: Segmenting and generating the invisible. In: CVPR (2018)
16. 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)
17. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Monocular dynamic view synthesis: A reality check. Advances in Neural Information Processing Systems **35**, 33768–33780 (2022)
18. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: ICCV (2023)
19. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
20. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. arXiv preprint arXiv:2311.18259 (2023)
21. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3749–3761 (2022)
22. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Haque, A., Tancik, M., Efros, A.A., Holynski, A., Kanazawa, A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. arXiv preprint arXiv:2303.12789 (2023)
24. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
26. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. ArXiv **abs/2204.03458** (2022), https://api.semanticscholar.org/CorpusID:248006185
27. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. arXiv preprint arXiv:2303.11989 (2023)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
29. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.: Ego-humans: An ego-centric 3d multi-human benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19807–19819 (2023)
30. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis

from multi-view video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5521–5531 (2022)
31. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
32. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: Dynibar: Neural dynamic image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4273–4284 (2023)
33. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
34. Ling, H., Kim, S.W., Torralba, A., Fidler, S., Kreis, K.: Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. arXiv preprint arXiv:2312.13763 (2023)
35. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
36. Liu, Y.L., Gao, C., Meuleman, A., Tseng, H.Y., Saraf, A., Kim, C., Chuang, Y.Y., Kopf, J., Huang, J.B.: Robust dynamic radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13–23 (2023)
37. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
38. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023)
39. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. arXiv preprint arXiv:2403.06351 (2024)
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
41. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
42. Ozguroglu, E., Liu, R., Surís, D., Chen, D., Dave, A., Tokmakov, P., Vondrick, C.: pix2gestalt: Amodal segmentation by synthesizing wholes. In: CVPR (2024)
43. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: NeRFies: Deformable neural radiance fields. ICCV (2021)
44. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
45. Po, R., Wetzstein, G.: Compositional 3d scene generation using locally conditioned diffusion. arXiv preprint arXiv:2303.12218 (2023)
46. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
47. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural radiance fields for dynamic scenes. In: CVPR (2021)

48. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
50. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., et al.: Infinite photorealistic worlds using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12630–12641 (2023)
51. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., et al.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21783–21794 (2024)
52. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single real image. arXiv preprint arXiv:2310.17994 (2023)
53. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
54. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
55. Shamsian, A., Kleinfeld, O., Globerson, A., Chechik, G.: Learning object permanence from video. In: ECCV (2020)
56. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. In: The Twelfth International Conference on Learning Representations (2023)
57. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
58. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., et al.: Text-to-4d dynamic scene generation. arXiv preprint arXiv:2301.11280 (2023)
59. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
60. Tokmakov, P., Jabri, A., Li, J., Gaidon, A.: Object permanence emerges in a random walk along memory. In: ICML (2022)
61. Tokmakov, P., Li, J., Burgard, W., Gaidon, A.: Learning to track with object permanence. In: ICCV (2021)
62. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12959–12970 (2021)
63. Tschernezki, V., Darkhalil, A., Zhu, Z., Fouhey, D., Laina, I., Larlus, D., Damen, D., Vedaldi, A.: Epic fields: Marrying 3d geometry and video understanding. arXiv preprint arXiv:2306.08731 (2023)

64. Van Hoorick, B., Tendulkar, P., Suris, D., Park, D., Stent, S., Vondrick, C.: Revealing occlusions with 4d neural fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3011–3021 (2022)
65. Van Hoorick, B., Tokmakov, P., Stent, S., Li, J., Vondrick, C.: Tracking through containers and occluders in the wild. In: CVPR (2023)
66. 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. arXiv preprint arXiv:2403.12008 (2024)
67. Walke, H., Black, K., Lee, A., Kim, M.J., Du, M., Zheng, C., Zhao, T., Hansen-Estruch, P., Vuong, Q., He, A., Myers, V., Fang, K., Finn, C., Levine, S.: Bridgedata v2: A dataset for robot learning at scale. In: Conference on Robot Learning (CoRL) (2023)
68. Wang, C., Eckart, B., Lucey, S., Gallo, O.: Neural trajectory fields for dynamic novel view synthesis. arXiv preprint arXiv:2105.05994 (2021)
69. Wang, C., Zhuang, P., Siarohin, A., Cao, J., Qian, G., Lee, H.Y., Tulyakov, S.: Diffusion priors for dynamic view synthesis from monocular videos. arXiv preprint arXiv:2401.05583 (2024)
70. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
71. Wang, L., Zhang, J., Liu, X., Zhao, F., Zhang, Y., Zhang, Y., Wu, M., Yu, J., Xu, L.: Fourier plenoctrees for dynamic radiance field rendering in real-time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13524–13534 (2022)
72. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
73. Weissenborn, D., Täckström, O., Uszkoreit, J.: Scaling autoregressive video models. ICLR (2020)
74. Wikipedia contributors: Camera dolly — Wikipedia, the free encyclopedia (2024), `https://en.wikipedia.org/wiki/Camera_dolly`, [Online; accessed 2024]
75. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Xinggang, W.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
76. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. arXiv preprint arXiv:2312.02981 (2023)
77. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: CVPR (2021)
78. Xiao, L., Nouri, S., Hegland, J., Garcia, A.G., Lanman, D.: Neuralpassthrough: Learned real-time view synthesis for vr. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–9 (2022)
79. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
80. Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., Sridhar, S.: Neural fields in visual computing and beyond. In: Computer Graphics Forum. vol. 41, pp. 641–676. Wiley Online Library (2022)

81. 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)
82. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning (2020)
83. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., et al.: Language model beats diffusion–tokenizer is key to visual generation. ICLR (2024)
84. Zhan, X., Pan, X., Dai, B., Liu, Z., Lin, D., Loy, C.C.: Self-supervised scene deocclusion. In: CVPR (2020)
85. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics (TOG) **40**(4), 1–18 (2021)
86. Zhang, Q., Wang, C., Siarohin, A., Zhuang, P., Xu, Y., Yang, C., Lin, D., Zhou, B., Tulyakov, S., Lee, H.Y.: Scenewiz3d: Towards text-guided 3d scene composition. arXiv preprint arXiv:2312.08885 (2023)
87. Zhao, X., Colburn, R.A., Ma, F., Bautista, M.Á., Susskind, J.M., Schwing, A.: Pseudo-generalized dynamic view synthesis from a video. In: The Twelfth International Conference on Learning Representations (2024)
88. Zhao, Y., Yan, Z., Xie, E., Hong, L., Li, Z., Lee, G.H.: Animate124: Animating one image to 4d dynamic scene. arXiv preprint arXiv:2311.14603 (2023)
89. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
90. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19855–19865 (2023)