LaVR: Scene Latent Conditioned Generative Video Trajectory Re-Rendering using Large 4D Reconstruction Models

Mingyang Xie[1,2] Numair Khan[1] Tianfu Wang[2] Naina Dhingra[1] Seonghyeon Nam[1] Haitao Yang[1]
Zhuo Hui[1] Christopher Metzler[2] Andrea Vedaldi[1,3] Hamed Pirsiavash[1,4] Lei Luo[1]

[1]Meta Reality Labs [2] University of Maryland [3]University of Oxford [4]UC Davis

https://lavr-4d-scene-rerender.github.io/

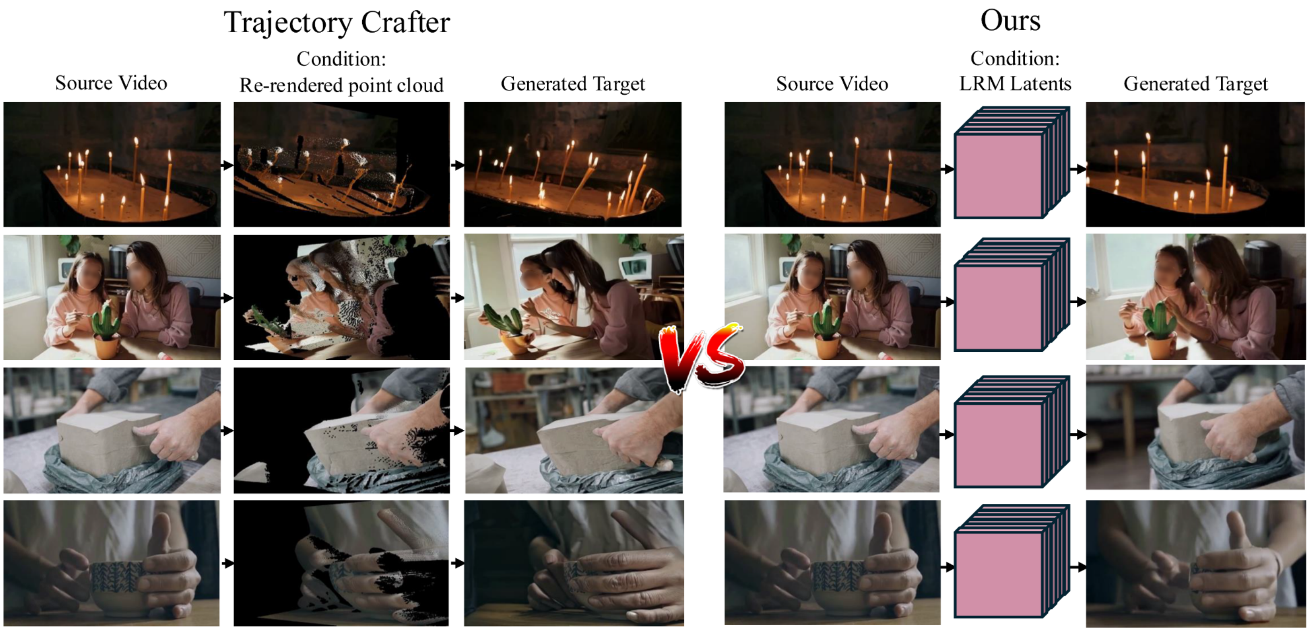

Figure 1. Our method addresses the problem of rendering geometrically consistent novel trajectories from a monocular source video. We propose to utilize the geometric knowledge of a pre-trained large reconstruction model (LRM) by conditioning the trajectory generation process on the latent state of a 4D LRM. Compared to prior methods which are conditioned on error-prone point cloud re-renderings of the source video, our method achieves state-of-the-art visual quality while maintaining a high level of geometric fidelity to the original scene.

Abstract

Given a monocular video, the goal of video re-rendering is to generate views of the scene from a novel camera trajectory. Existing methods face two distinct challenges. Geometrically unconditioned models lack spatial awareness, leading to drift and deformation under viewpoint changes. On the other hand, geometrically-conditioned models depend on estimated depth and explicit reconstruction, making them susceptible to depth inaccuracies and calibration errors. We propose to address these challenges by using the implicit geometric knowledge embedded in the latent space of a large 4D reconstruction model to condition the video generation process. These latents capture scene structure in a continuous space without explicit reconstruction. Therefore, they provide a flexible representation that allows the pretrained diffusion prior to regularize errors more effectively. By jointly conditioning on these latents and source camera poses, we demonstrate that our model achieves state-of-the-art results on the video re-rendering task.

1. Introduction

Video re-rendering, or novel trajectory synthesis, aims to visualize a dynamic scene from new and unseen camera paths. Unlike standard video generation, this task requires modeling both scene dynamics and underlying geometry to maintain temporal and spatial coherence under arbitrary camera motion. Controllable trajectory synthesis enables various applications including re-rendering captured scenes for film, and generating immersive experience from a single video. However, the problem is inherently challenging. Monocular inputs provide weak geometric supervision, forcing models to jointly infer structure, motion, and lighting.

Existing works approach this problem from two directions. Geometrically-conditioned methods [9, 24, 40] model scene geometry using point clouds or meshes and re-render novel views. Although physically grounded, these methods [24, 40] rely on accurate depth estimation, so any errors propagate into the re-rendered point clouds and cause shape distortions in novel views, where objects appear stretched or compressed along the depth direction and parallax becomes inconsistent. To address this issue, geometrically unconditioned methods [4] have been recently proposed to avoid explicit conditioning, and instead generate videos using only the input video and target trajectory. Such methods achieve strong visual realism–largely inherited from the pretrained video diffusion prior—-but struggle with geometric consistency across viewpoints. This motivates a new approach that can capture the strengths of both directions by providing geometric guidance without depending on precise depth.

We propose a model for novel trajectory synthesis that conditions a video diffusion backbone on a latent 4D scene representation extracted from monocular videos. Instead of explicit 4D conditioning, we encode input videos into a latent space that captures appearance, geometry, and dynamics, enabling the model to follow novel camera trajectories while maintaining coherent structure and parallax.

This formulation is enabled by recent large 4D reconstruction models [15, 30, 32] (LRMs), which have shown that a feed-forward network can extract rich latent representations from monocular frames and decode them into depth, pose, or approximate novel views. These models demonstrate that implicit geometry structure can be captured without explicit optimization or volumetric reconstruction, providing exactly the type of geometry-aware cues our framework leverages.

While both point clouds and LRM-produced latents can provide geometric cues, they condition the generative process in fundamentally different ways. Point-cloud pipelines reconstruct geometry from estimated depth and re-render it from the target viewpoint, so any depth error directly manifests as distorted shapes, incorrect parallax, or missing regions—acting as a rigid geometric constraint that leaves the generative model little flexibility to correct mistakes. In contrast, implicit geometry latents provide structural guidance

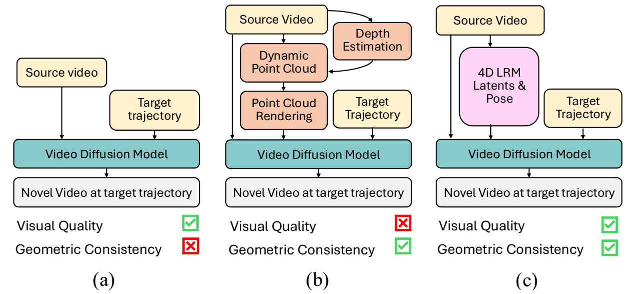

Figure 2. **Architecture comparison. (a)** Unconditioned methods for novel trajectory generation achieve high visual quality but lack geometric awareness, leading to inconsistencies. **(b)** Conditioning on 4D point cloud renders provides consistency, but reduces quality as the depth estimation and point cloud generation stages are sensitive to errors. **(c)** Our proposed architecture utilizes the implicit geometric knowledge of a pre-trained large 4D reconstruction model (LRM) to achieve both high quality and consistency.

in a softer, non–pixel-aligned form. Because the diffusion model is pretrained on large-scale video data with strong priors over plausible motion and scene structure, it can regularize small geometric inconsistencies in these latents. This combination yields geometric cues that are both informative and robust to depth noise, motivating our design choice.

As a result, our latent space conditioning formulation provides strong geometric priors without relying on accurate depth estimation. It enables the model to maintain stable parallax and coherent structure under large viewpoint changes, producing geometrically consistent videos along arbitrary camera trajectories. Fig. 1 illustrates the qualitative advantages of our approach compared with a state-of-the-art explicitly 4D-conditioned method [40], and Fig. 2 compares our overall paradigm with existing baselines. In summary, our novel contributions are:

- We propose to use the latent state of a large reconstruction models to provide geometric conditioning without relying on explicit depth or point-cloud reconstruction.
- We present a lightweight adapter module that compresses and integrates the latents from a state-of-the-art 4D reconstruction model with VAE-encoded video latents for efficient consumption by a pre-trained diffusion backbone.
- We conduct extensive evaluation to show that our approach outperforms both geometrically-conditioned and unconditioned baselines on quantitative and qualitative metrics.

2. Related Work

Unsupervised 3D/4D Scene Reconstruction. Neural scene representations such as NeRF [5, 6, 19] and 3D Gaussian Splatting [14] have advanced novel view synthesis by reconstructing scene geometry from posed images. While effective for static scenes with sufficient multi-view coverage, their

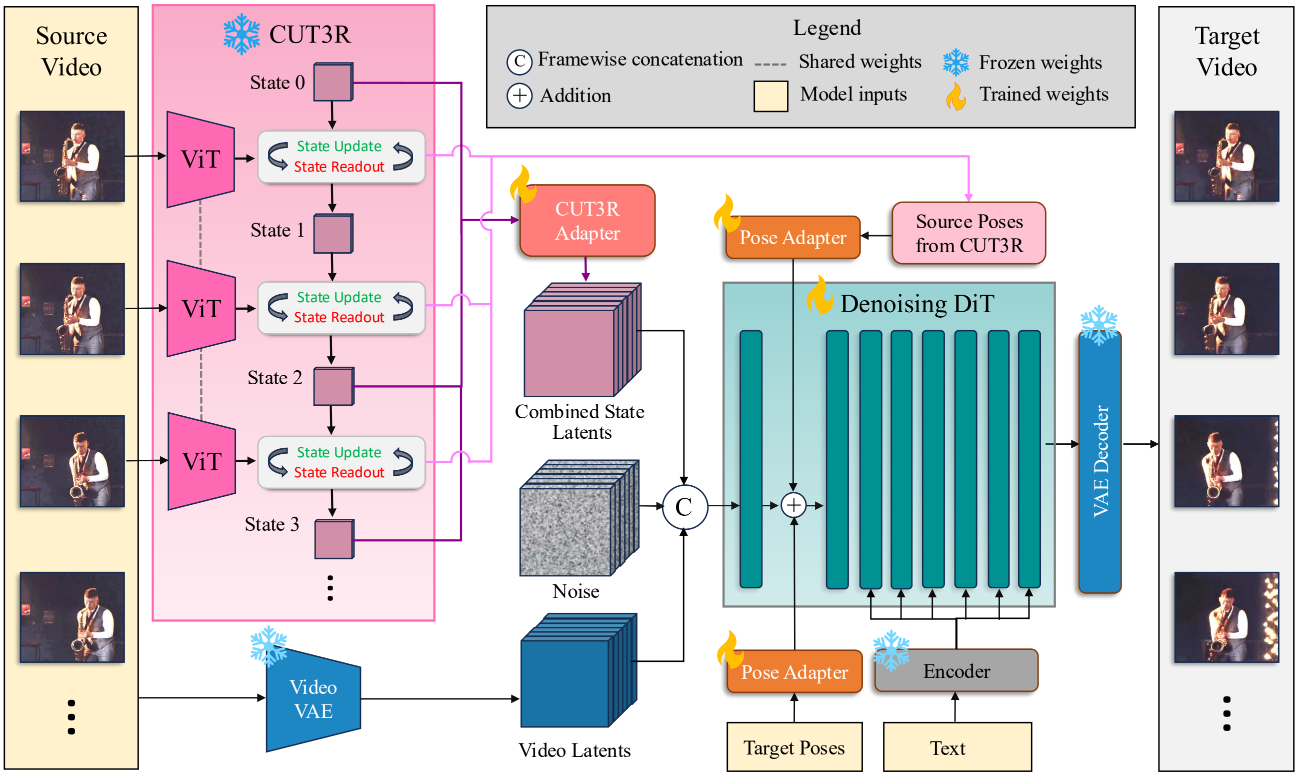

Figure 3. **Pipeline overview.** Given a monocular source video, our method generates a novel video of the same scene at a target camera trajectory using a video diffusion model. To ensure geometric consistency, we condition the model on latents from CUT3R [32], a pre-trained 4D reconstruction model. We use four signals from the source video: the standard video VAE latents, CUT3R's 4D latents, source camera poses, and an encoded text description of the scene. A novel adapter architecture aligns the CUT3R and VAE latents, and allows these to be fed to the model in a computationally feasible manner. The source camera poses come from CUT3R, and are added to the DiT's intermediate activations after passing through a small MLP-based adapter. Another MLP processes the target poses at which the novel video is rendered. Note that only the projection and self-attention layers of the DiT are trainable, other parameters are frozen.

reliance on accurate and complete geometry makes them difficult to apply when only monocular video frames are available. Subsequent work has extended these approaches to dynamic scenes [8, 21–23, 25, 31, 35]. Early monocular methods rely on depth-based warping [38] with later improvements in occlusion handling, while newer techniques adopt neural dynamic representations [21–23, 25] or Gaussian Splatting variants [8, 31, 35] with additional regularization to stabilize time-varying geometry. Nonetheless, these pipelines still often degrade when the target camera trajectory departs significantly from the observed views.

Large 3D/4D Reconstruction Models. Large 3D and 4D reconstruction models [29, 30, 32, 33, 42] leverage high-capacity architectures and large-scale pretraining to estimate scene structure from single images, image pairs, or short video clips. Early feed-forward approaches such as DUST3R [33] demonstrate that transformer-based correspondence aggregation alone can recover camera poses and dense geometry without iterative optimization. Building on this idea, more recent systems such as SPANN3R [29], CUT3R [32], MONST3R [42], and MegaSAM [15] scale model capacity and training data to achieve more generalizable reconstructions, and capture time-varying or 4D scene structure. These large reconstruction models provide strong geometric priors that can serve as effective conditioning signals for generative and camera-controlled video synthesis.

Generative Novel View/Trajectory Synthesis. Recent works in this direction have explored using generative models to achieve controllable camera trajectories for video synthesis [1–3, 10, 11, 11, 12, 18, 20, 24, 26, 28, 34, 37, 39, 41]. These methods typically condition video diffusion models on camera poses or trajectory signals to guide viewpoint changes. In the dynamic setting, TrajectoryCrafter [40], Gen3C [24], and EX-4D [12] generate new views by conditioning on rendered point clouds or meshes, offering strong consistency but inheriting depth-related brittleness. By contrast, ReCamMaster [4] directly synthesizes videos along new trajectories without explicit 4D structure, providing greater flexibility at the cost of weaker geometric stability.

Explicit 4D pipelines therefore excel at enforcing geometry but are vulnerable to reconstruction errors, whereas non-4D-conditioned methods remain more robust but struggle to maintain spatial coherence under large camera motion.

3. Method

Given a source video, our goal is to synthesize novel frames along a user-specified camera trajectory while preserving the scene content and dynamics of the source. To achieve this, we propose conditioning a video-to-video diffusion model on the latent state of a large 4D reconstruction model (LRM).

Our proposed approach exploits the fact that the latent state aggregates scene structure and camera motion without explicit geometric reconstruction. Thus, it circumvents the error-prone approach of depth estimation and point-cloud/mesh reconstruction adopted by existing models with geometric conditioning [24, 36, 40]. In addition, these prior works utilize geometry only indirectly through renderings. These renderings – which suffer from distortions and holes – bake the reconstructed scene into a 2D image, and leave the generative model limited room to reason about the underlying geometry and correct errors. In contrast, the latent space of an LRM encodes geometry and camera poses in a high-dimensional feature space. Not only does this preserve the entire 4D scene structure, the continuous representation allows more flexibility for the pre-trained video diffusion prior to regularize inconsistencies during the generation stage.

In addition to the latent state, we use the source camera poses and a text prompt as secondary conditions to provide additional context for the input frames. The user-specified target poses are supplied as a control signal to steer the denoising process toward the desired camera path.

3.1. Scene State Latents with CUT3R

We use CUT3R [32] as a representative 4D LRM for our task. CUT3R provides consistent feed-forward reconstruction from a monocular video by maintaining a persistent latent state that aggregates multi-view information over time and reflects the evolving 3D understanding of the scene. Multiple heads decode the state to recover the pose, world-space pointmaps, and depth for each frame of the input video, thereby demonstrating that the latent state contains strong geometry and motion cues for downstream conditioning.

The latent state is represented as a set of s tokens $\{\ell^i \in \mathbb{R}^d\}_{i=1}^s$ which are updated at each time step by the ViT-encoded frame of the source video (Figure 3). We use the state tokens $\mathcal{S} = \{\{\ell_t^i \in \mathbb{R}^d\}_{i=1}^s, t = 1, 2, ..., T\}$ for all T frames of the video in order to preserve temporal changes in both scene content and camera pose.

3.2. Adapting CUT3R Latents to DiT Backbone

While the CUT3R state latents $\mathcal{S}$ provide rich geometric cues, they are not directly compatible with the inputs of the diffusion backbone. State-of-the-art video diffusion models learn

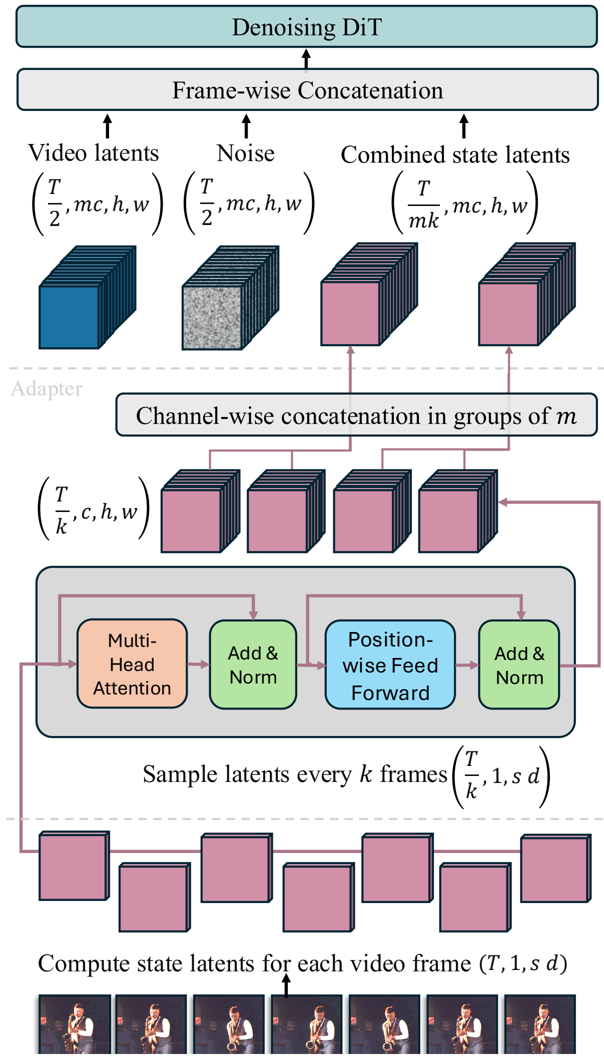

Figure 4. **Proposed CUT3R Adapter.** Our lightweight adapter compresses CUT3R's per-frame latent tokens into geometry-aware features that align with the representation used by the diffusion model. The shape of features at each stage is shown in brackets.

to denoise a VAE-encoded latent representation of the source frames, which then passes through a decoder to produce the RGB output. Thus, we introduce a lightweight adapter to compress CUT3R's latent tokens into geometry-aware features aligned with the latent representation expected by the denoising diffusion transformer (DiT).

Preserving the backbone's learned temporal inductive biases requires that the CUT3R and VAE latents be aligned along the frame dimension. This is because the DiT is trained to process frame-indexed tokens via temporal self-attention, and maintaining the structure of each token is important to ensure the injected geometry cues do not disrupt pre-trained behavior. Channel-wise concatenation would alter the shape of each temporal token, and require substantial retraining.

However, directly concatenating and feeding all CUT3R tokens into the diffusion model is computationally infeasible, as self-attention scales quadratically with sequence length. We found that using the full CUT3R latent sequence, a single training iteration could take two minutes on 8×H200 GPUs!

Therefore, our proposed CUT3R adapter performs temporal and token-level compression before injecting geometry into the diffusion backbone (Figure 4). More specifically, starting with a tensor of shape $(T, 1, s, d)$ representing the CUT3R state latent $\mathcal{S}$, we sub-sample every k-th frame to reduce the sequence length to $\frac{T}{k}$. The sub-sampled tokens are then passed through a two-layer transformer decoder with 12 heads per layer, which fuses information within each frame yielding a tensor sized $(\frac{T}{k}, c, h, w)$ where (h, w) is the size of the video VAE latent in the spatial dimension. We temporally aggregate these features by merging every m consecutive frames into a single feature map. We choose m and c such that $m \cdot c$ equals the channel size of the video VAE latent. Thus, we obtain $\frac{T}{mk}$ CUT3R latent groups. Each group of shape $(m{\cdot}c,\ h,\ w)$ summarizes the geometric information within a local temporal window.

We concatenate these groups with the VAE-encoded source video latents and the noisy latents along the frame dimension, preserving the temporal token layout expected by the pretrained DiT and avoiding any modification to the backbone architecture.

3.3. Training Strategy

We train the CUT3R and pose adapters from scratch, and finetune a subset of DiT layers. These latter include the DiT projector, and all self-attention blocks. The remaining layers of the DiT, and the Video VAE are frozen to preserve pretrained priors. We train the model on the synthetic MultiCamVideo dataset from ReCamMaster [4] which provides multiple posed trajectories for dynamic scenes. We randomly select two trajectories per scene as source and target.

We use a standard conditional flow-matching loss [17] as the training objective. Specifically, given the clean *target* latent z_0, a noise sample $\epsilon \sim \mathcal{N}(0, I)$, and a diffusion timestep $t \sim \mathcal{U}(0, 1)$, we set the interpolated latent $z_t = (1 - t)z_0 + t\epsilon$. The DiT predicts the velocity field conditioned on the adapted CUT3R state latents Z_c and the source video latents Z_s using the following objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\, z_0,\, \epsilon}\left[\left\|v_\theta(z_t,\ t,\ Z_c,\ Z_s)\ -\ (\epsilon - z_0)\right\|_2^2\right].$$

To allow fast convergence of the geometry-conditioned pathway and preserve pretrained priors, we use a 3× higher learning rate for the CUT3R adapter than other components.

4. Experiments

Baselines. We compare our results with three baseline methods for novel trajectory synthesis from monocular videos:

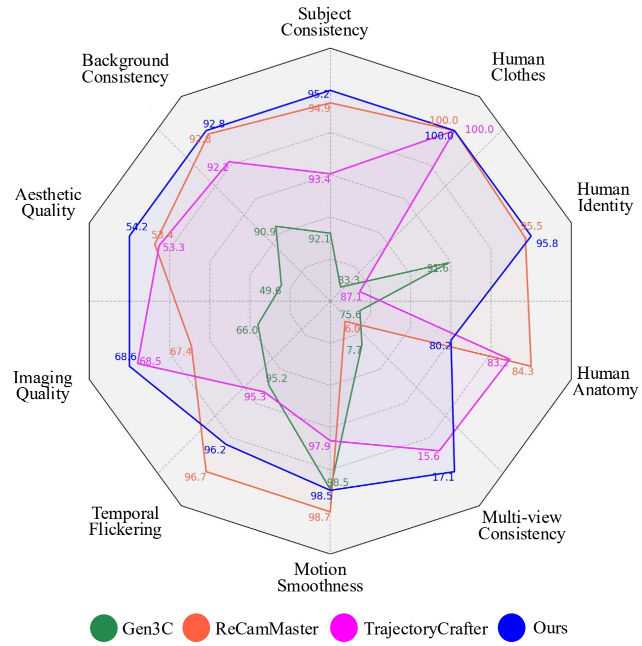

Figure 5. **Evaluation on the VBench [13, 43] suite of metrics.** We highlight relative differences by normalizing each metric over all baselines. Our method shows all-around high performance, achieving the best results for multi-view, subject, and background consistency.

Gen3C [24], TrajectoryCrafter [40], and ReCamMaster [4]. The first two use a re-rendered point cloud as a conditioning signal, while the last is un-conditioned. For each baseline, we use the code implementation and pretrained models provided by the authors. We set camera parameters to ensure all evaluated methods follow similar trajectories.

Implementation Details. We train our model on 8×H200 GPUs for 15K iterations with a batch size of eight. We used a learning rate of 6×10^{-5} for the CUT3R adapter, and 2×10^{-5} for the other trainable parameters.

Evaluation Dataset. To create a diverse evaluation dataset, we obtain a hundred dynamic-scene videos from Pexels (a large stock-footage platform), and fifty static-scene videos from DL3DV [16]. All videos are resampled to 33 frames and resized to a resolution of 480×832. For each video, we evaluate all methods under four different novel camera trajectories. To ensure a fair comparison, each video is accompanied by exactly the same text caption.

4.1. Results

Qualitative results for all baselines and our method are presented in Figures 6, and 9. Gen3C and TrajectoryCrafter frequently suffer from unnatural warping due to errors in the reconstructed point clouds. This is clearly illustrated in Figure 7 which shows the rendered point cloud condition along

Figure 6. **Qualitative evaluation of novel views.** We compare frames from new camera trajectories rendered by redirecting the source video. Both Gen3C [24] and TrajectoryCrafter [40] are conditioned on re-rendered point clouds, and suffer from unnatural warping artifacts. ReCamMaster [4] is not geometrically conditioned, and hallucinates implausible content in unseen regions (*top row:* interior of tent; *middle row*: cat's tail; *bottom row:* haystacks in the background). Compared to the baselines, our results look natural and geometrically consistent.

	Cycle Consistency			VBench Consistency		
Method	PSNR↑	LPIPS↓	CLIP↑	Subject↑	Multi-view↑	Background↑
Gen3C [24]	20.62	23.23	97.47	92.07	7.695	90.91
TrajectoryCrafter [40]	14.84	41.59	95.05	93.38	15.57	92.21
ReCamMaster [4]	17.75	32.63	97.03	94.95	5.975	92.76
Ours	**20.74**	**22.47**	**98.07**	**95.22**	**17.11**	**92.83**

Table 1. **Quantitative comparison of consistency**. We highlight the metrics in blue, proportional to their percentile. The values for LPIPS, CLIP, and all VBench metrics are $\times 10^{-2}$. Our method shows strong performance on all metrics, achieving the best results on cycle consistency. Please refer to the qualitative results in Figures 8 and 9, and to the supplementary material for video results.

	Pose Reconstruction Error		
Method	Abs($\mathbf{t}$) ↓	Rel($\mathbf{t}$) ↓	Rel($\mathbf{R}$) ↓
Gen3C	24.45	12.00	0.641
TrajectoryCrafter	16.53	10.52	0.442
ReCamMaster	21.83	12.43	0.518
Ours	**14.39**	**7.798**	**0.411**

Table 2. **Target pose reconstruction accuracy**. We evaluate the absolute $\mathrm{Abs}(\cdot)$ and relative $\mathrm{Rel}(\cdot)$ errors in camera translation $\mathbf{t}$ (in millimeters) and rotation $\mathbf{R}$ (in degrees). While we achieve consistently high rotational and translational accuracy, the unconditioned ReCamMaster [4] fails to follow the target trajectory closely. We highlight the metrics in blue, proportional to their percentile.

k	m	Cycle PSNR↑	Multi-view↑	Abs($\mathbf{t}$) ↓	Rel($\mathbf{t}$) ↓	Rel($\mathbf{R}$) ↓
1	8	19.24	14.43	16.80	9.620	0.420
2	4	19.50	**17.52**	16.84	8.374	0.430
4	2	**20.74**	17.11	**14.39**	**7.798**	**0.411**
8	1	19.15	16.13	16.65	8.917	0.453

Table 3. **Ablation on CUT3R Adapter Parameters.** The combination of m and k for all 4 rows corresponds to the same number of CUTER latent groups to be fed into the DiT, implying the same amount of DiT computation. $k = 4$ and $m = 2$ achieves the best overall performance. More details are in Fig. 4 and Sec. 4.2.

with the generated output of TrajectoryCrafter. ReCamMaster, on the other hand, lacking any geometric conditioning, hallucinates implausible content and fails to maintain object consistency across occlusions (Figure 8). Our method generates views that look more natural and consistent.

Pose Reconstruction Accuracy. We evaluate the accuracy of the generated trajectories by running the state-of-the-art dynamic scene bundle adjustment method of Chen*et al.* [7] on the output of each baseline. We then use the Umeyama algorithm [27] to align the predicted poses with the ground truth targets, and calculate average absolute and relative errors over all frames. We present the results in Table 2. Our

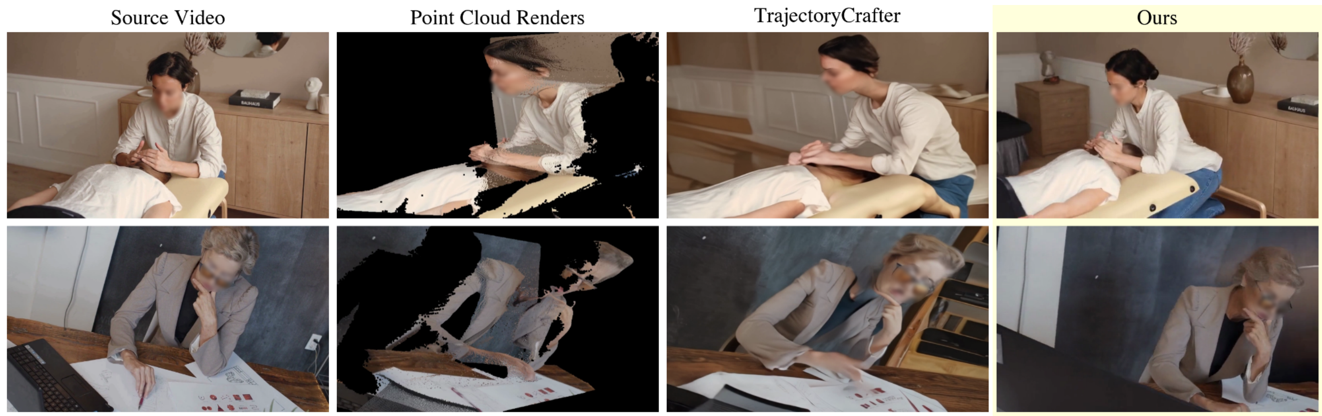

Figure 7. **Disadvantages of geometric conditioning via re-rendered point clouds.** Visualizing the point cloud renders of TrajectoryCrafter we see that depth scale ambiguity, empirically estimated intrinsics, and holes and misalignment errors in the point cloud can create warped conditioning images which lead to unnatural outputs. Our results do not suffer from such artifacts.

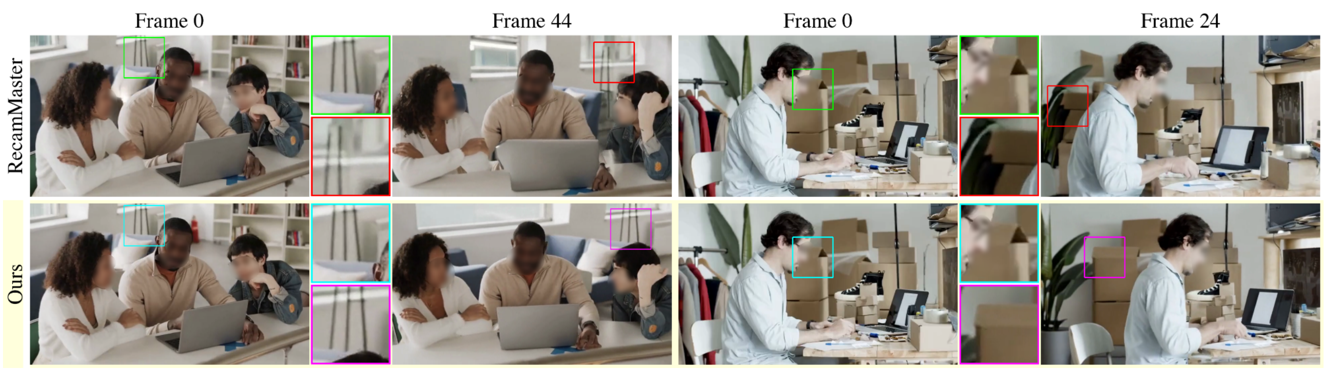

Figure 8. **Qualitative evaluation of geometric consistency across frames.** We show the beginning and end frames for two novel video trajectories from ReCamMaster [4], and our method. Lacking any geometric conditioning, ReCamMaster fails to maintain consistency across occlusions (*Left:* a leg of the lamp vanishes. *Right:* the cardboard box opens after reappearing behind the subject). The latent 4D condition of our method produces a more stable reconstruction. Please refer to the supplementary material for video results.

method closely follows the target trajectory, with the lowest translation and rotation errors of all baselines.

Cycle Consistency. To measure the consistency of the generated videos, we evaluate the symmetry between generated frames along a cyclical target trajectory. For static scenes, the generated views should match when the camera revisits the same pose. Thus, we evaluate all baselines for cycle consistency using 50 random videos from the DL3DV dataset. Our method outperforms TrajectoryCrafter and ReCamMaster, and has consistently lower error than Gen3C (Table 1).

Video Quality. We evaluate video generation quality using the VBench 1 & 2 [13, 43] suite of metrics on dynamic video inputs (Figure 5 and Table 1). Our method shows all-round high performance without any large failures, and achieves the best results on all consistency metrics.

4.2. Ablation Study

Recall in Figure 4 how we reduce the computation of the DiT self-attention layers by (1) temporally sampling one CUT3R latent every k frames and (2) grouping the resulting features into channel-wise concatenation groups of size m, yielding a smaller number of "CUT3R latent groups" to concatenate with the video latents along the frame dimension. To maintain reasonable training speed, we empirically fix the total number of CUT3R latent groups to 10. In addition, the VAE-encoded video latents have a fixed channel dimension of 16, which enforces the constraint $mc = 16$. Under these conditions, we perform an ablation study to determine the best relationship between m and k. Tab. 3 shows that the optimal configuration is achieved with $k = 4$ and $m = 2$.

Limitations Our method currently struggles with dynamic transparent objects (e.g., a glass cup being lifted by a person), likely due to CUT3R's limited ability to estimate reliable

Figure 9. **Qualitative evaluation of novel trajectories across frames.** Baseline methods conditioned on re-rendered point clouds suffer from unnatural **stretching artifacts, and missing details** (Row 1, Col. 4; Row 3, Col. 4). The unconditioned baseline **hallucinates a third arm** (Row 2, Col. 4). We avoid these pitfalls by using the latent state of a pre-trained 4D reconstruction model as a *soft* geometric condition.

geometry for such dynamic transparent materials.

5. Conclusion

We present a method for novel camera trajectory synthesis in dynamic scenes by conditioning a video diffusion model on latents from a large 4D reconstruction models (LRM). These latents provide geometry-aware guidance in a 'soft' form, allowing the pretrained diffusion prior to regularize local inconsistencies and avoid the errors and rigidity of rendered point clouds. Experiments on both static and dynamic scenes demonstrate that our approach achieves stronger geometric consistency, and higher visual quality than existing geometry-conditioned and unconditioned baselines methods.

6. Acknowledgement

We sincerely thank Bo Zhu and Denis Demandolx for their support and valuable discussions that improved this work.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 3

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025.

[3] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *arXiv preprint arXiv:2412.07760*, 2024. 3

[4] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 2, 3, 5, 6, 7, 1

[5] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[6] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2

[7] Weirong Chen, Ganlin Zhang, Felix Wimbauer, Rui Wang, Nikita Araslanov, Andrea Vedaldi, and Daniel Cremers. Back on track: Bundle adjustment for dynamic scene reconstruction. *arXiv preprint arXiv:2504.14516*, 2025. 6

[8] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[9] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025. 2

[10] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3

[11] Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3

[12] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh, 2025. 3

[13] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5, 7

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 2

[15] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[16] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5

[17] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ArXiv*, abs/2210.02747, 2022. 5

[18] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 3

[19] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf. *Communications of the ACM*, 65:99 – 106, 2020. 2

[20] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulo, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10258–10268, 2024. 3

[21] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 3

[22] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020.

[23] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 3

[24] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 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. 2, 3, 4, 5, 6

[25] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 3

[26] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 3

[27] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002. 6

[28] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*, pages 313–331. Springer, 2024. 3

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

[30] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[31] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[32] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2, 3, 4

[33] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. *Computer Vision and Pattern Recognition*, 2023. 3

[34] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[35] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 3

[36] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory, 2025. 4

[37] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3

[38] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. 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*, pages 5336–5345, 2020. 3

[39] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024. 3

[40] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025. 2, 3, 4, 5, 6

[41] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3, 1

[42] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3

[43] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 5, 7

LaVR: Scene Latent Conditioned Generative Video Trajectory Re-Rendering using Large 4D Reconstruction Models

Supplementary Material

7. Webpage

Please refer to the webpage in our supplementary materials for more video results, including comparisons with baselines.

8. Training Dataset

We train the model on the synthetic MultiCamVideo dataset from ReCamMaster [4] which contains 136K dynamic scenes created using Unreal Engine 5. For each dynamic scene, it provides 10 synchronized videos with *randomly generated* camera trajectories. We randomly select two trajectories per scene as source and target.

9. Model Size

Both our model and ReCamMaster [4] have approximately 1.3B parameters, while TrajectoryCrafter [41] and Gen3C have approximately 5B and 7B parameters, respectively. Despite our model's relatively small size, we still achieve the overall best performance on novel trajectory synthesis.