GEO4D: Leveraging Video Generators for Geometric 4D Scene Reconstruction

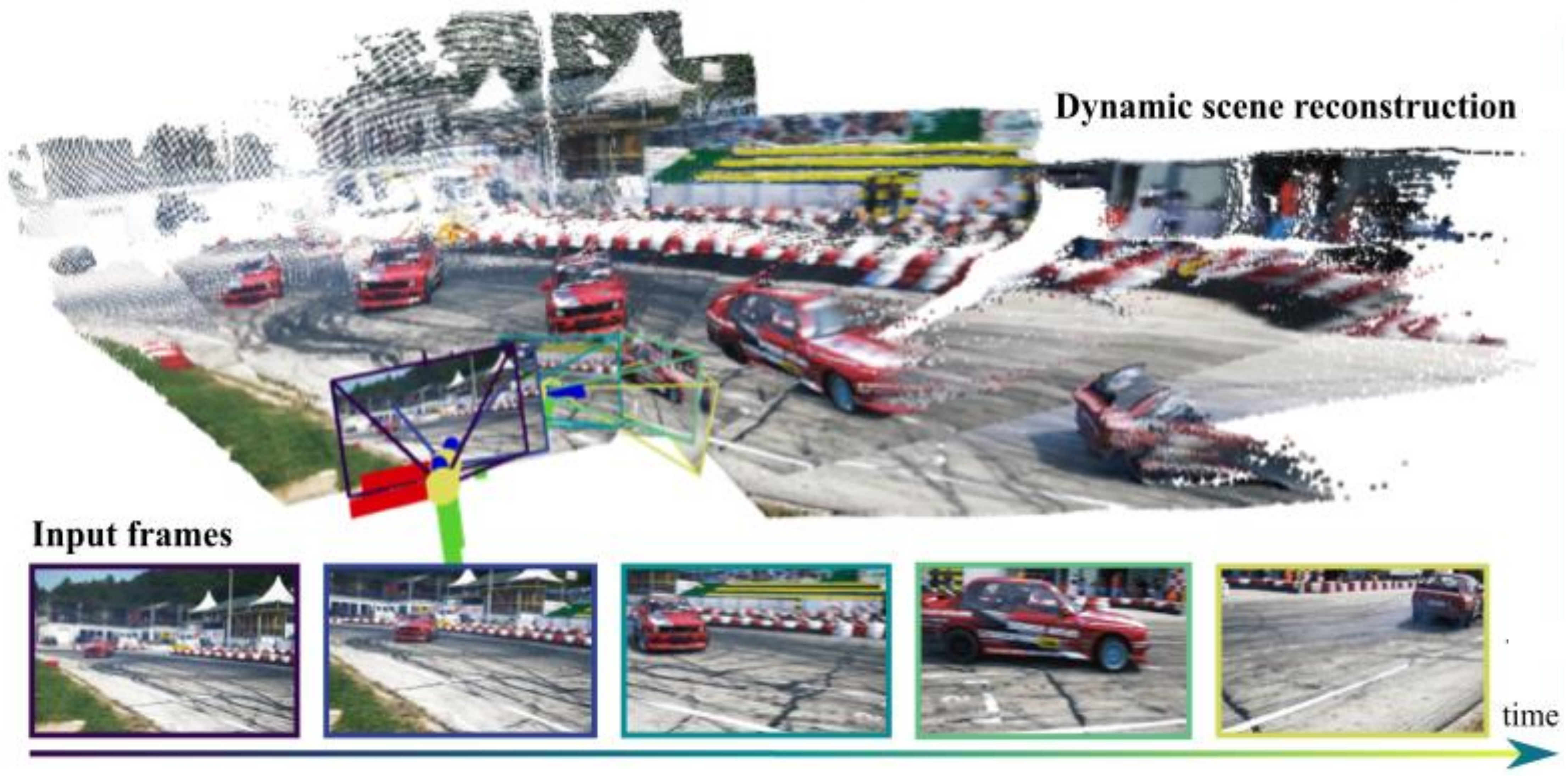
Zeren Jiang[1], Chuanxia Zheng[1], Iro Laina[1], Diane Larlus[2], Andrea Vedaldi[1]
[1]Visual Geometry Group, University of Oxford [2]Naver Labs Europe

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Motivation

Goal: Reconstruct dynamic scenes from monocular in-the-wild RGB videos

Challenges: Limited 4D data with ground-truth geometric annotations for training



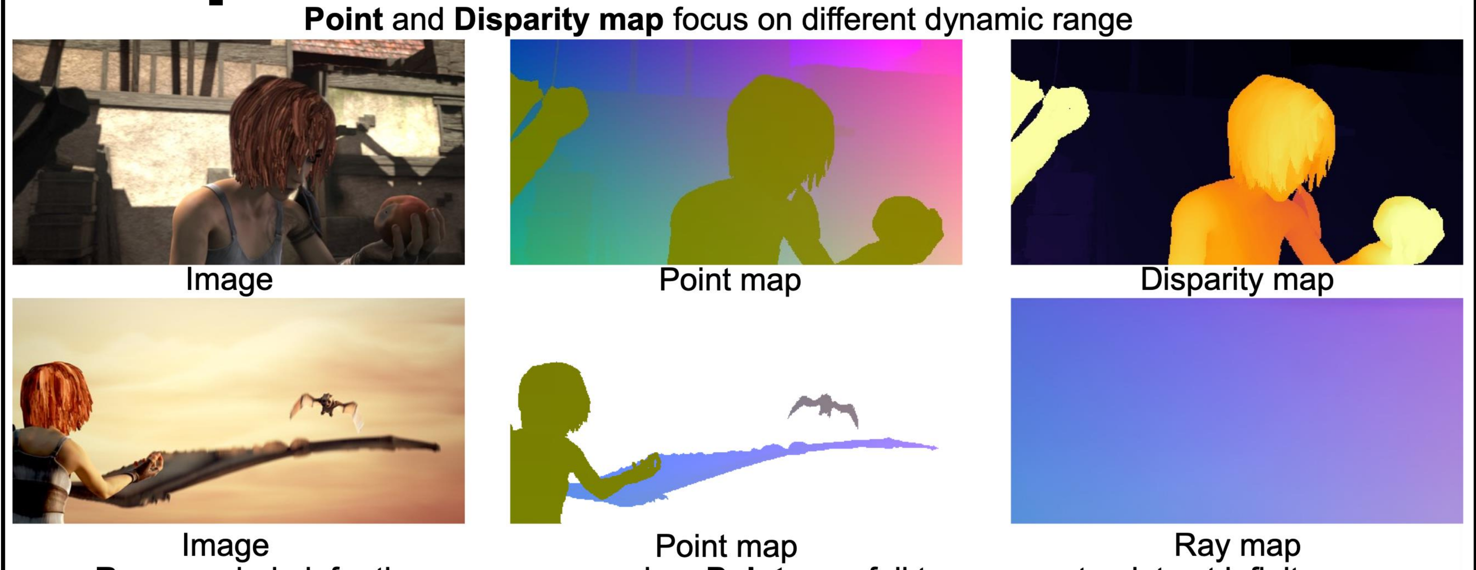
Solution: Repurposing video diffusion model for 4D scene reconstruction

Contribution

- Show that pre-trained **video generators** work well as **priors for 4D reconstruction**
- **A multi-modal geometric representation** that helps the video diffusion model to learn consistent geometry during training.
- **A lightweight multi-modal alignment** that fuses partially redundant geometric modalities at test time for coherent and robust 4D reconstruction.

Representation

Point and **Disparity map** focus on different dynamic range



Ray map help infer the camera pose when **Point map** fail to represent points at infinity

Evaluation

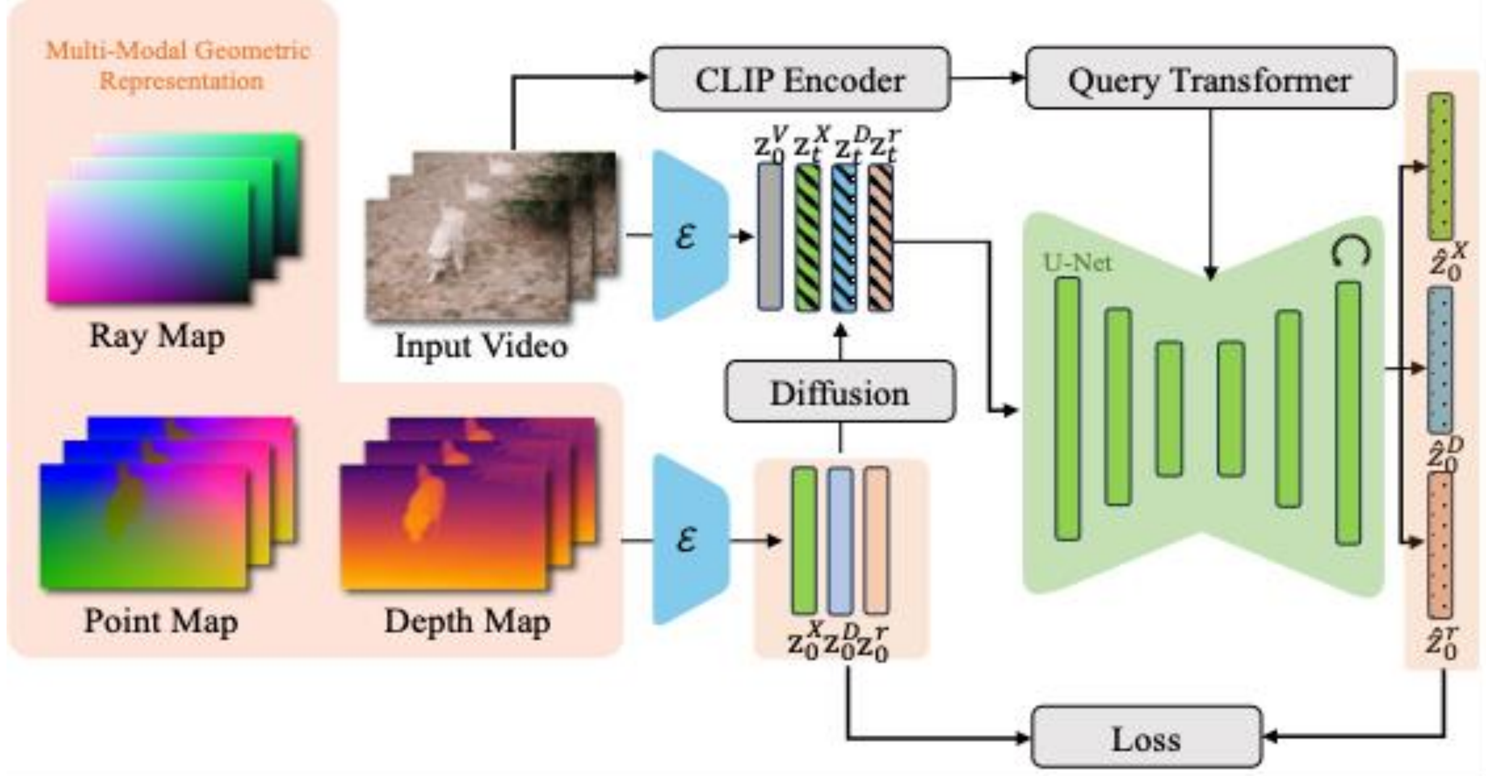
Category	Method	Sintel		Bonn		KITTI	
		Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
Single-frame depth	Marigold	0.532	51.5	0.091	93.1	0.149	79.6
	Depth-Anything-V2	0.367	55.4	0.106	92.1	0.140	80.4
Video depth	NVDS	0.408	48.3	0.167	76.6	0.253	58.8
	ChronoDepth	0.687	48.6	0.100	91.1	0.167	75.9
	DepthCrafter*	0.270	69.7	0.071	**97.2**	0.104	89.6
Video depth & Camera pose	Robust-CVD	0.703	47.8	-	-	-	-
	CasualSAM	0.387	54.7	0.169	73.7	0.246	62.2
	MonST3R	0.335	58.5	0.063	96.4	0.104	89.5
	Ours	**0.205**	**73.5**	**0.059**	**97.2**	**0.086**	**93.7**

Video depth estimation on Sintel, Bonn, and KITTI datasets.

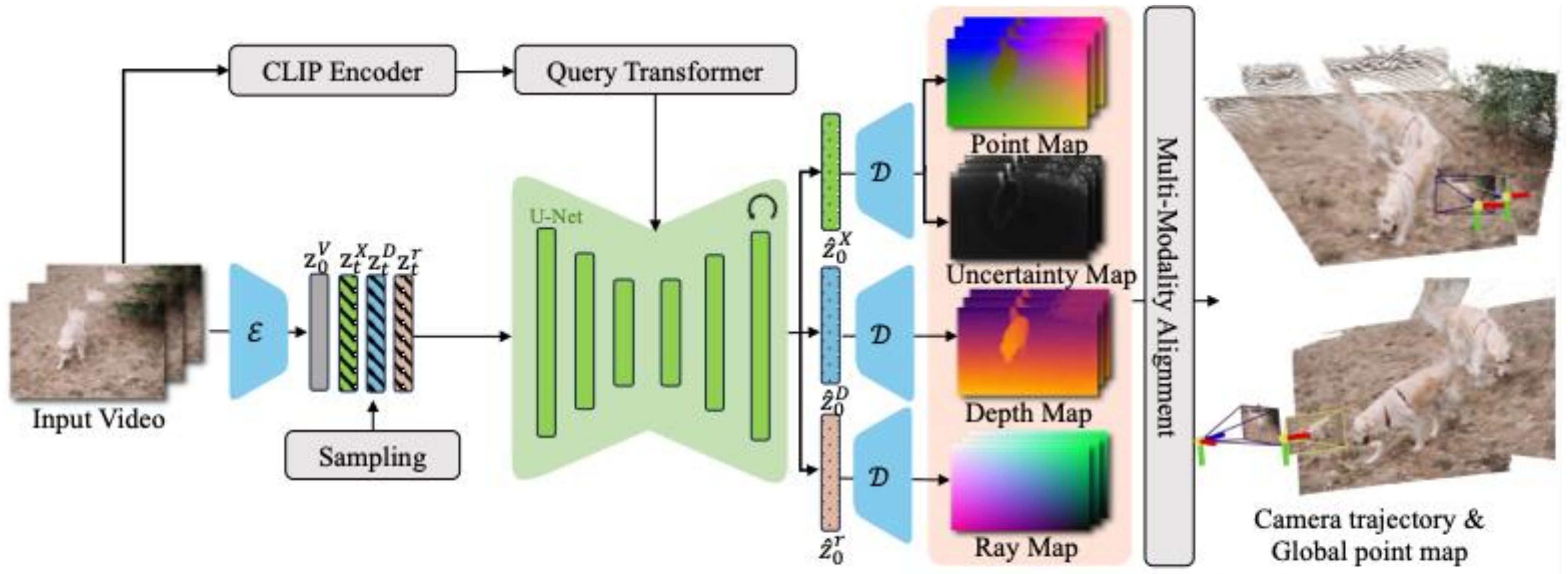
Training			Inference			Video Depth		Camera Pose		
Point Map	Depth Map	Ray Map	Point Map	Depth Map	Ray Map	Abs Rel ↓	$\delta < 1.25$ ↑	ATE ↓	RPE trans ↓	RPE rot ↓
✓	-	-	✓	-	-	0.232	71.3	0.335	0.076	0.731
✓	✓	✓	✓	-	-	0.223	72.5	0.237	0.070	0.566
✓	✓	✓	-	✓	-	0.211	73.4	-	-	-
✓	✓	✓	-	-	✓	-	-	0.268	0.192	1.476
✓	✓	✓	✓	✓	✓	**0.205**	**73.5**	**0.185**	**0.063**	**0.547**

Ablation study for the different modalities of the geometric representation.

Method

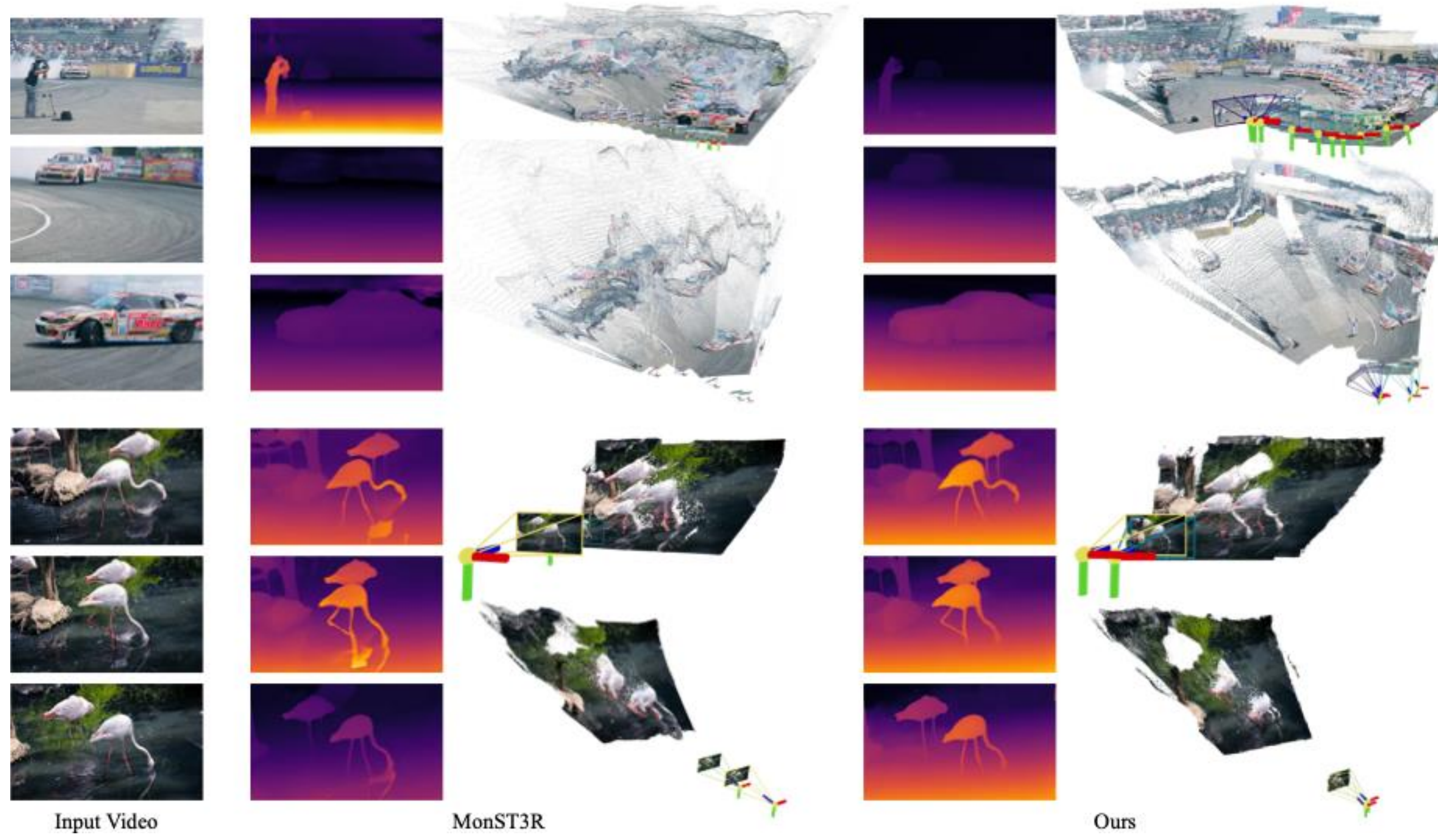


We cast 4D video reconstruction as conditional generation of 4D latent features $\mathbf{z}_t^{\mathrm{X}}, \mathbf{z}_t^{\mathrm{D}}, \mathbf{z}_t^{\mathrm{r}}$ from an RGB video encoder adapted for point, depth and camera ray encoding. The input video is injected as condition via cross-attention in the denoising U-Net, after and a query transformer.



During **inference**, iteratively denoised latent features $\widehat{\mathbf{z}_0^{\mathrm{X}}}, \widehat{\mathbf{z}_0^{\mathrm{D}}}, \widehat{\mathbf{z}_0^{\mathrm{r}}}$ are decoded by the fine-tuned VAE decoder, followed by multi-modal alignment optimization for coherent 4D reconstruction.

Visualization



- Group-wise inference allow **fast motion** to be reconstructed consistently
- Prior geometry knowledge from the video generator enables accurate reconstruction under **deceptive cues**, (e.g., reflections)

Project Page: geo4d.github.io