



Stereo World Model

Yang-Tian Sun¹, Zehuan Huang², Yifan Niu², Lin Ma³, Yan-Pei Cao², Yuewen Ma³, Xiaojuan Qi¹

¹The University of Hong Kong, ²VAST, ³ByteDance Pico



Abstract

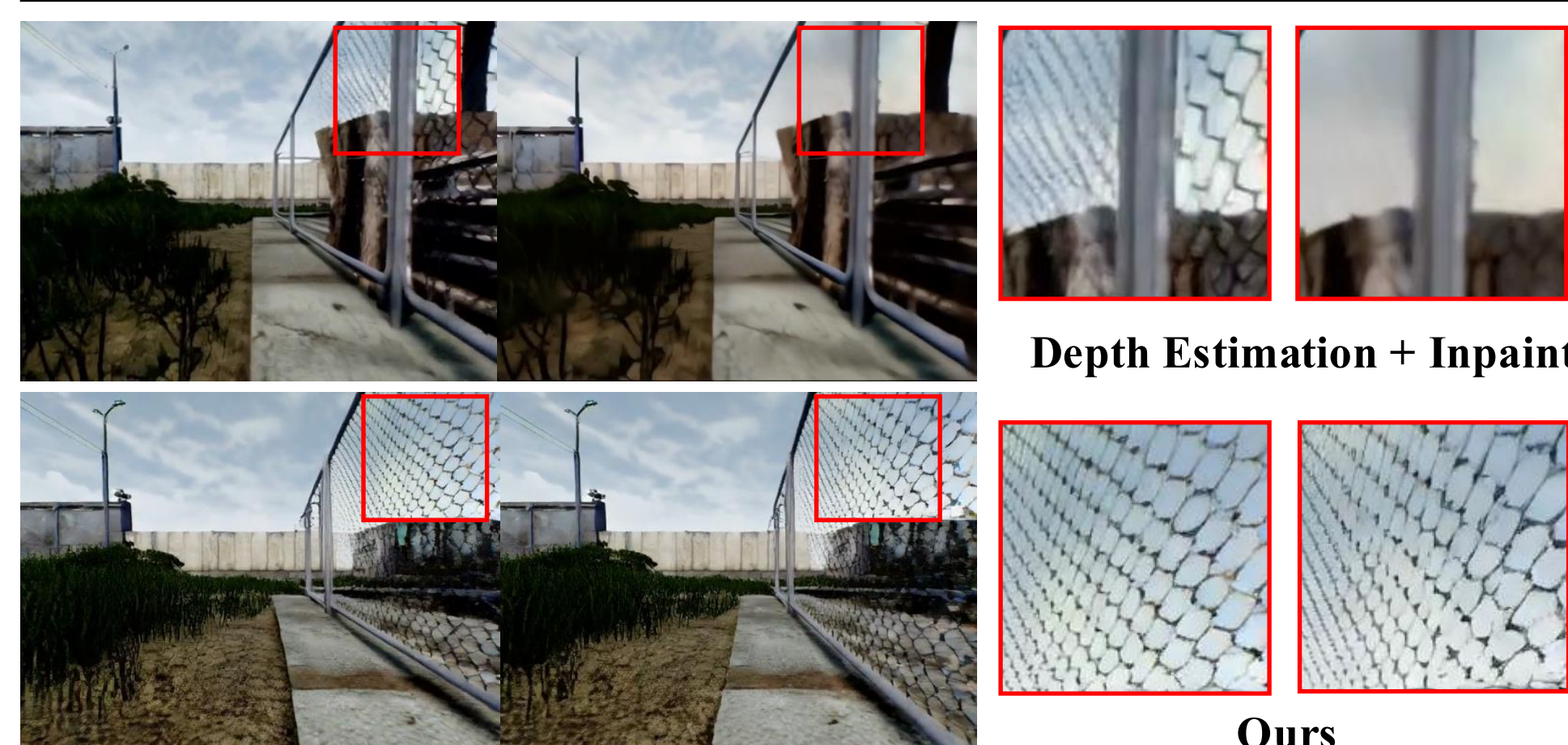
We present **StereoWorld**, a camera-conditioned stereo world model that jointly learns appearance and binocular geometry for end-to-end stereo video generation.

Our approach introduces two key designs:

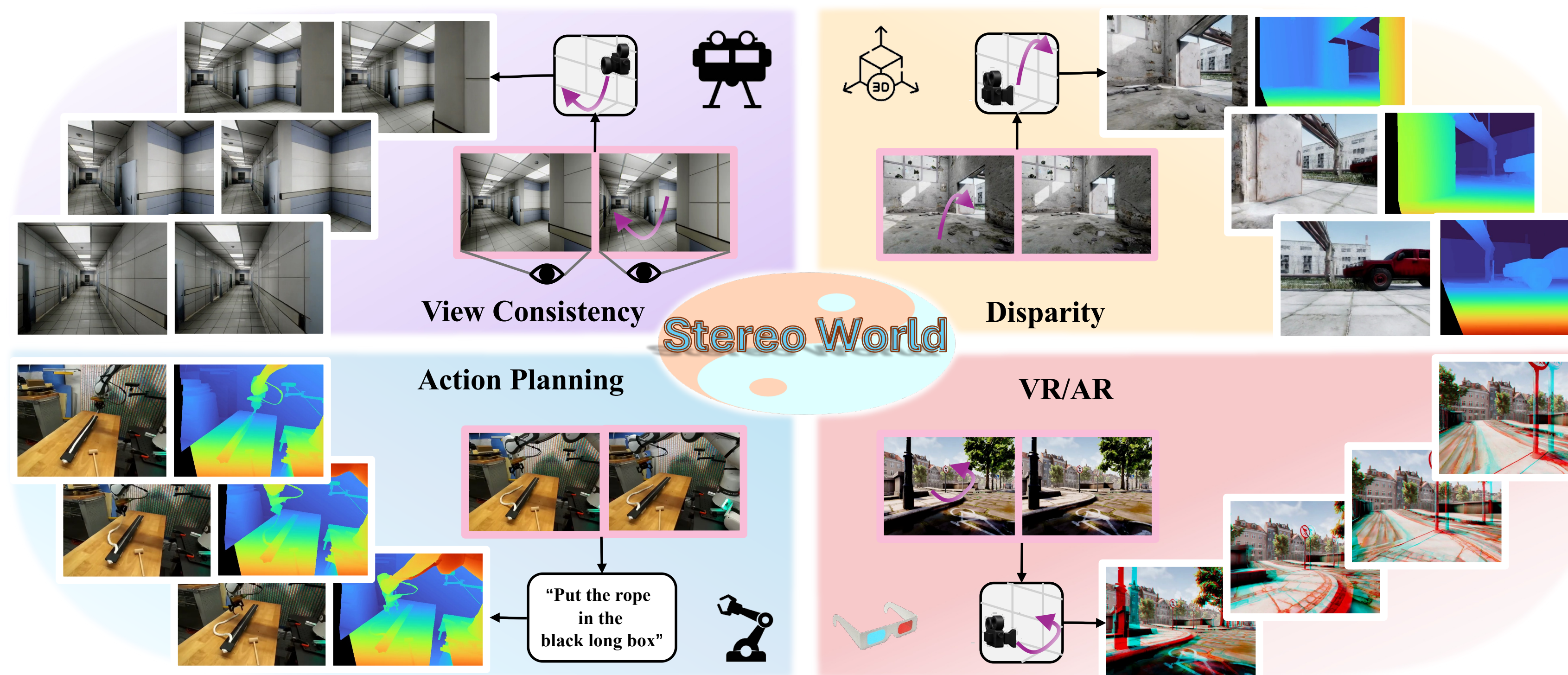
- (1) a **unified camera-frame RoPE** that augments latent tokens with camera-aware rotary positional encoding, enabling relative, view- and time-consistent conditioning while preserving pretrained video priors via a stable attention initialization;
- (2) a **stereo-aware attention** decomposition that factors full 4D attention into 3D intra-view attention plus horizontal row attention, leveraging the epipolar prior to capture disparity-aligned correspondences with substantially lower compute.

Motivation

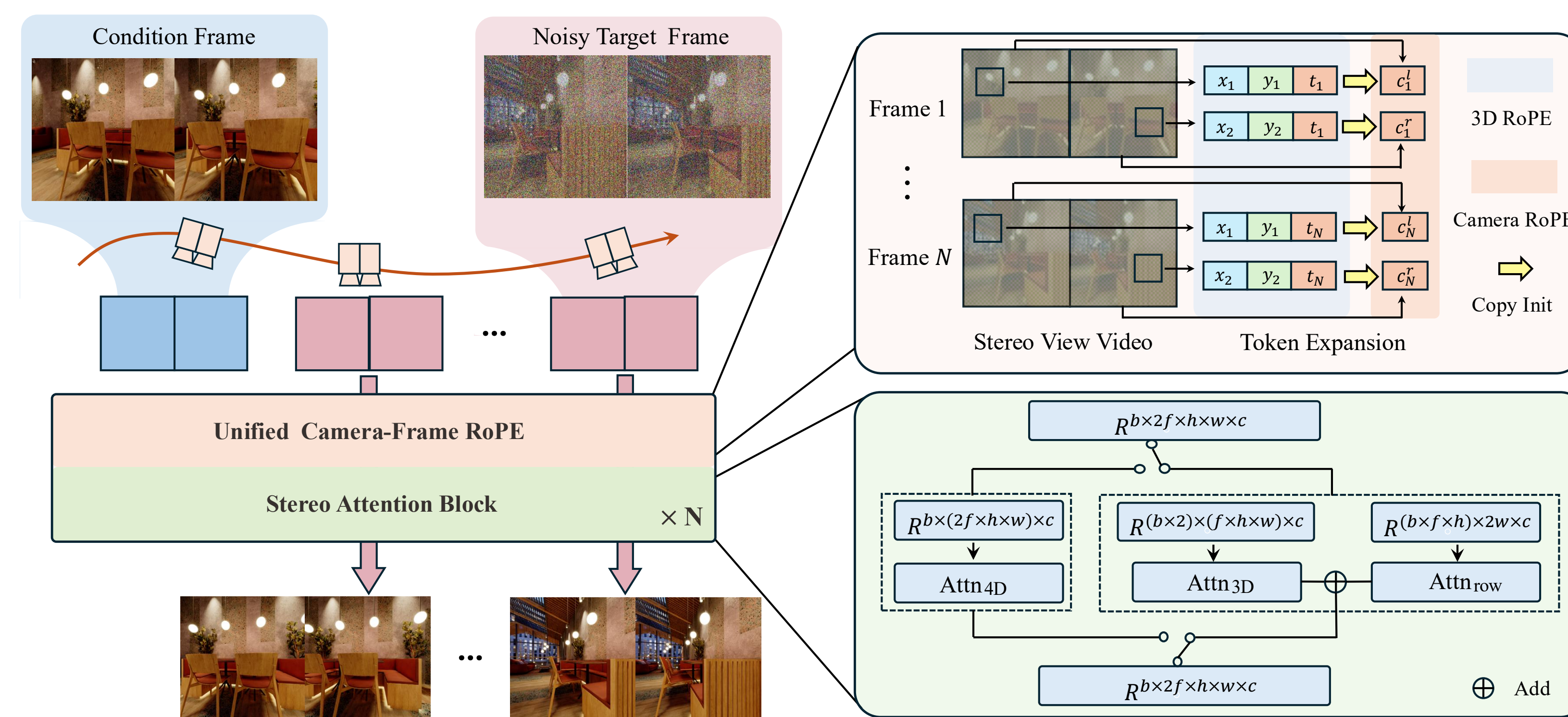
World Model Modality	Monocular, RGB	Monocular, RGBD	Stereoscopy, RGB
Geometry	✗	Relative, Limited Range	Metric
Same Domain with Pre-Trained Model	✓	✗	✓
VR Display	Rely on External Depth Estimation and Inpainting Model		End-to-End



- **Biologically Inspired:** stereo vision is the dominant perceptual mechanism in biological systems provides direct, robust geometric cues to 3D scene structure
- **End-to-End:** Avoids explicit depth estimation and inpainting by directly generating stereo videos
- **Prior Preserved:** generate frames in the same domain with pre-trained video models



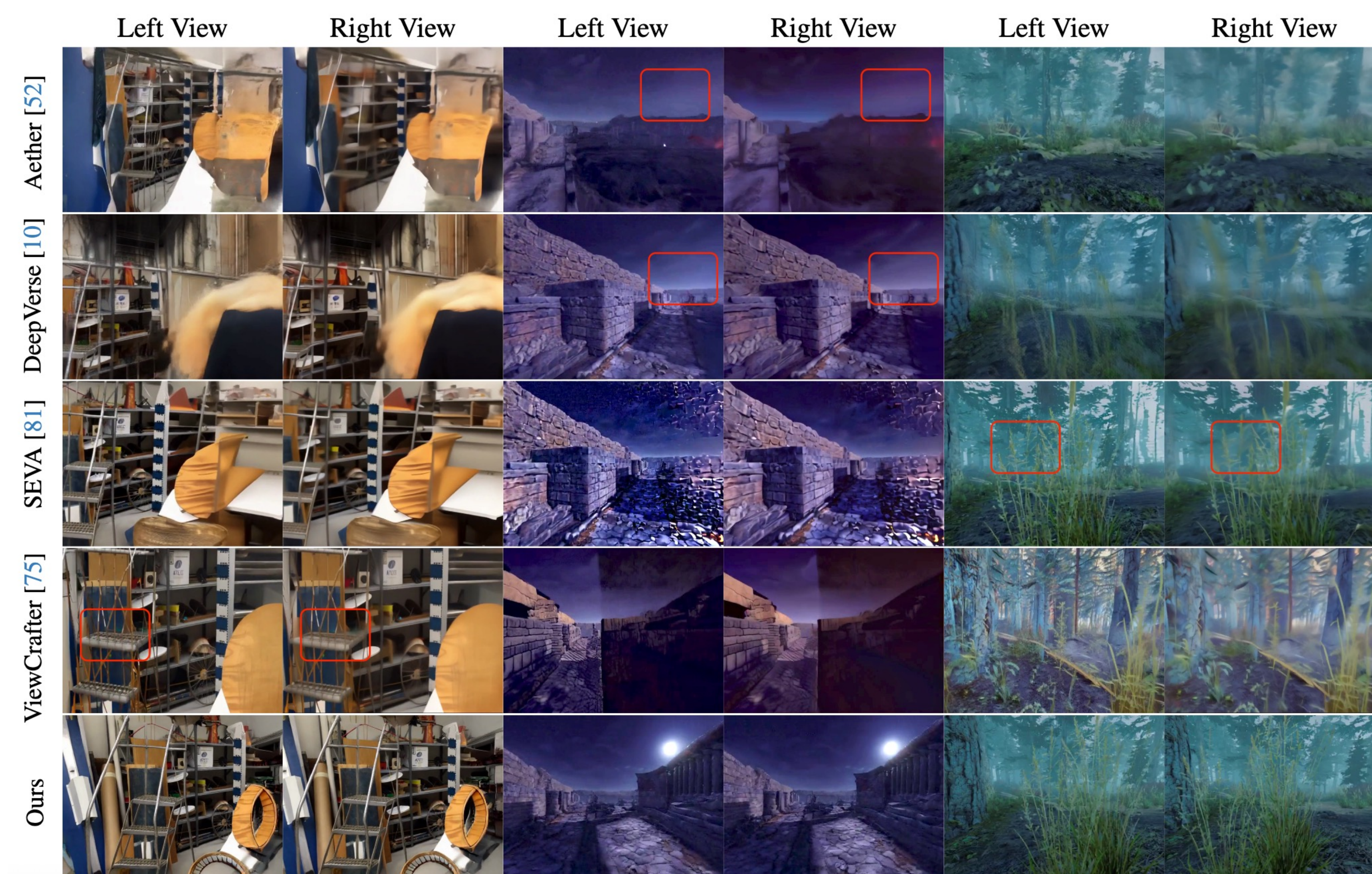
Method



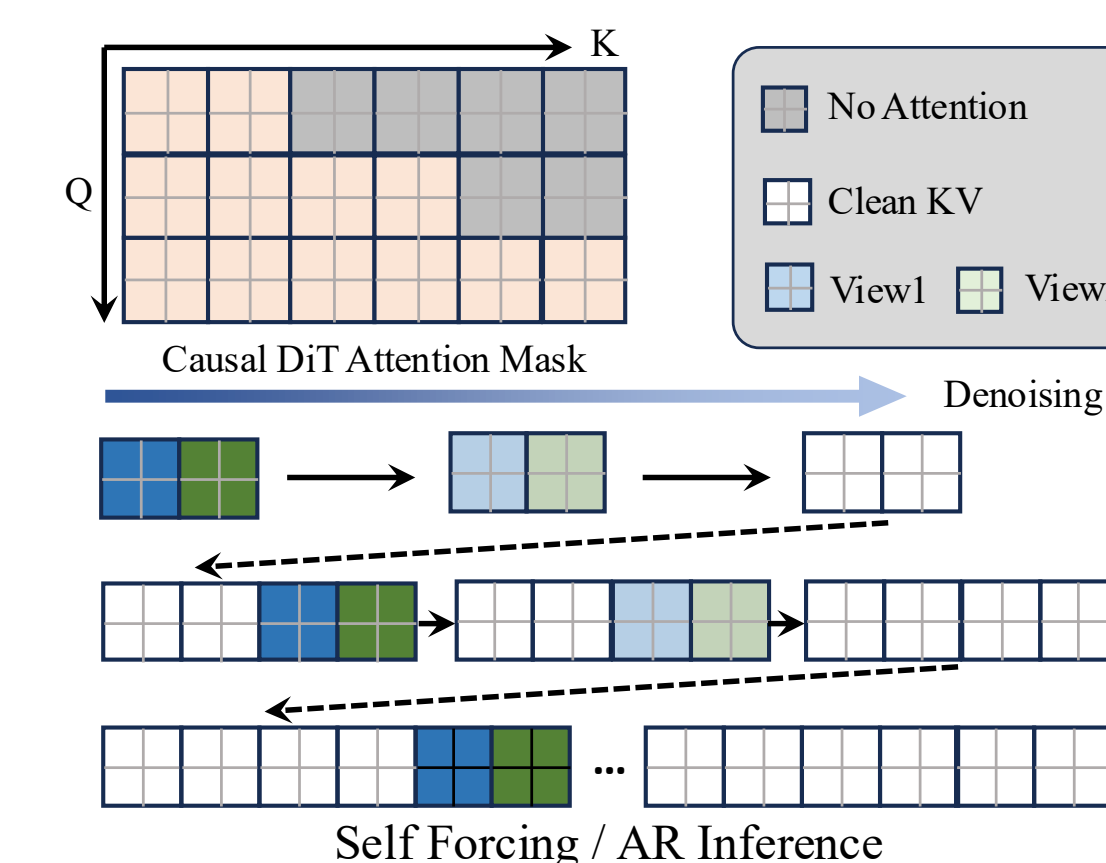
- **Unified Camera-Frame RoPE:** inject camera positional encodings both temporally and spatially, minimally perturbing pretrained priors
- **Stereo-Aware Attention:** decompose 4D spatial-temporal attention into (a) intra-view 3D attention and (b) cross-view attention

Result

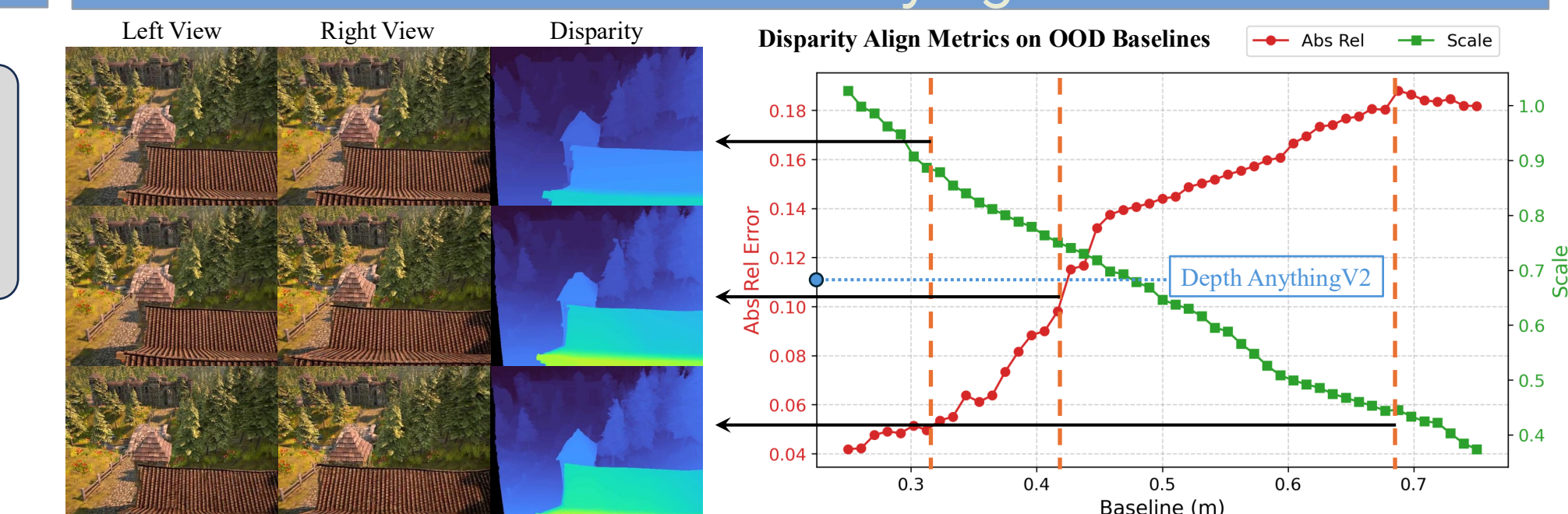
Method	Modality	Visual Quality				Camera Accuracy		View Synchronization			FPS↑
		FID↓	FVD↓	CLIP-T↑	CLIP-F↑	RotErr↓	TransErr↓	Mat. Pix.(K)↑	FVD-V↓	CLIP-V↑	
Voyager [26]	RGBD	226.97	170.37	24.85	97.03	1.34	0.25	4.26	55.45	91.41	0.03
DeepVerse [10]	RGBD	191.32	176.72	24.59	97.31	1.51	0.16	4.48	33.50	93.86	0.35
Aether [52]	RGBD	185.72	152.97	24.93	97.14	1.50	0.13	4.35	42.07	93.71	0.11
SEVA [81]	RGB	195.70	170.92	24.77	98.11	1.09	0.51	4.49	31.10	94.73	0.10
ViewCrafter [75]	RGB	211.89	185.76	25.02	96.15	1.24	0.20	4.49	42.10	93.51	0.13
Ours Monocular	RGB	126.83	96.87	24.97	97.12	1.36	0.14	✗	✗	✗	✗
Ours Stereo	RGB	111.36	83.04	25.74	97.55	1.01	0.11	4.56	22.00	97.50	0.49



Long-Video Distillation



Generalization to Varying Baselines



- StereoWorld handles video sequences with varying stereo baselines, despite being trained only on fixed-baseline data.