



DynamicVerse: A Physically-Aware Multimodal Framework for 4D World Modeling

Kairun Wen^{1*,†} Yuzhi Huang^{1*} Runyu Chen¹ Hui Zheng¹ Yunlong Lin¹ Panwang Pan¹ Chenxin Li² Wenyan Cong³ Jian Zhang¹ Junbin Lu⁴
Chenguo Lin⁵ Dilin Wang⁶ Zhicheng Yan⁶ Hongyu Xu⁶ Justin Theiss⁶ Yue Huang¹ Xinghao Ding^{1✉} Rakesh Ranjan⁶ Zhiwen Fan³

* Equal Contribution; † Project Leader; ✉ Corresponding Author

¹XMU ²CUHK ³UT Austin ⁴UW ⁵PKU ⁶Meta



Motivation

- Limited Data Diversity: indoor scenes, autonomous driving, and “sim-to-real” gaps.
- Lack of Physical Scale: Limited metric-scale geometry and detailed semantic captions.
- Scalability Issues: Multiple sensor-based methods are not scalable.

Contribution

- DynamicGen: automated data engine, which generates physically-aware multi-modal 4D data
- DynamicVerse: a large-scale 4D dataset featuring diverse dynamic scenes accompanied by rich multi-modal annotations including metric-scale point maps, camera parameters, object masks with corresponding categories, and detailed descriptive captions.



Object Category: Running Lady

Camera Caption: The camera moves forward, following the running lady from behind, resulting in slightly unsteady motion with shaking. As the video concludes, the camera stops tracking, tilts up, and pans left to reveal the scene ahead.

Object Caption: An elderly lady with short white hair, wearing a vibrant multicolored blouse and black pants, walks with a steady, rhythmic gait. Their arms are slightly bent, holding a small object. Maintaining an upright posture with head tilted forward, they move at a consistent pace, suggesting focus, purpose, or familiarity with their path.

Scene Caption: A lively scene inside a spacious, well-lit restaurant, characterized by wooden floors, large windows with natural light, and a mix of modern and rustic decor including exposed brick walls and furniture. An elderly lady walks purposefully through the warm, inviting space, her stride steady, bustling with diverse patrons seated enjoying meals or conversations at many tables set with cutlery and glasses. The ambiance is cozy yet sophisticated.

Figure 1. The overview of physically-aware multi-modal 4D world modeling framework DynamicVerse.

Performance: Metric-scale 4D Reconstruction

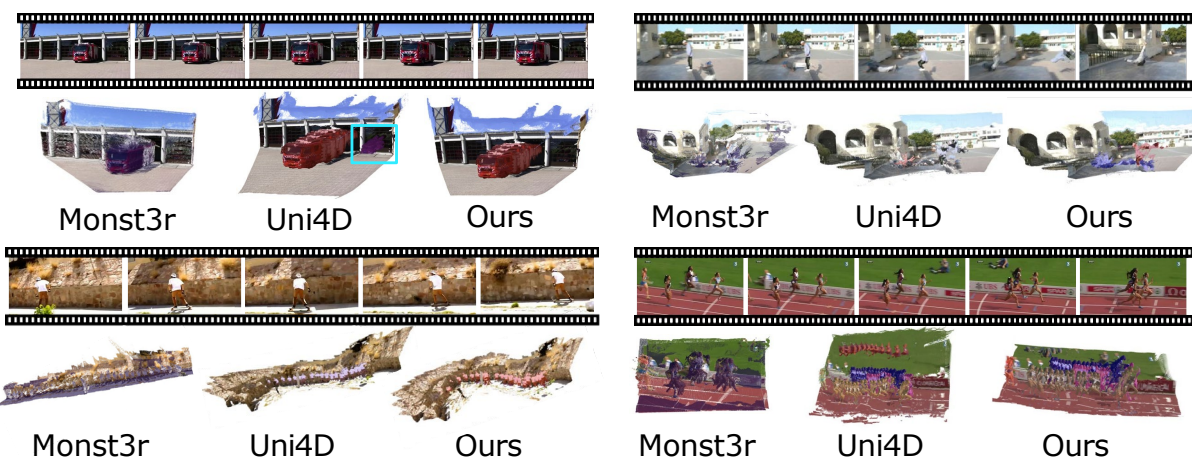


Figure 2. Visual comparisons of 4D reconstruction on in-the-wild data between baselines and DynamicGen.

Pipeline: DynamicGen

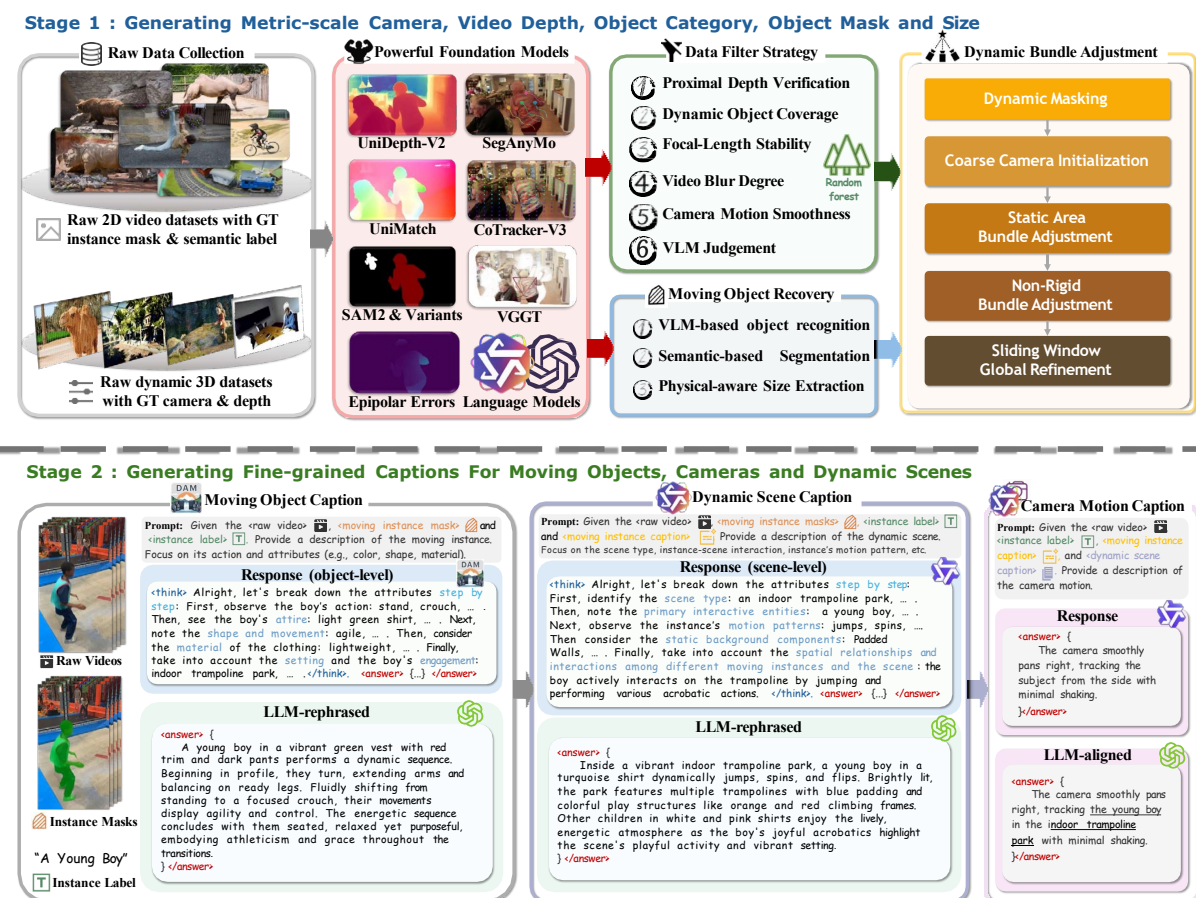


Figure 3. The physically-aware multi-modal 4D data generation pipeline DynamicGen.

Performance: Moving Object Segmentation

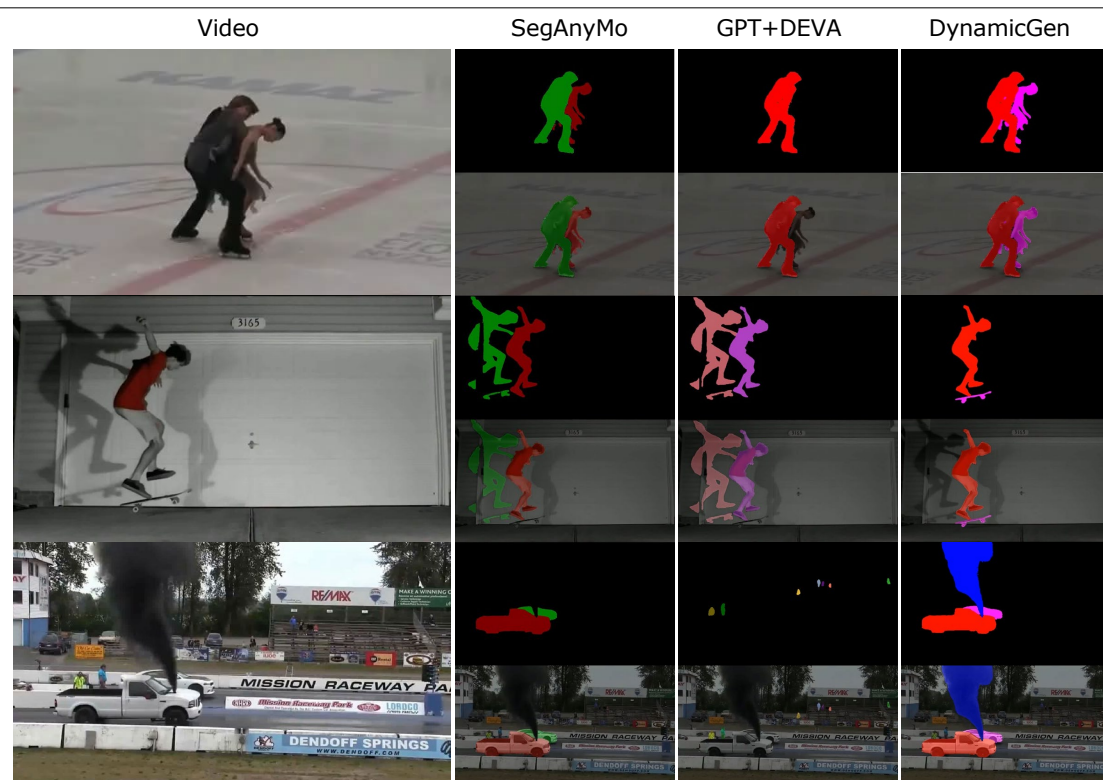


Figure 4. Visual comparisons of moving object segmentation between baselines and DynamicGen.

Dataset: DynamicVerse

Table 1. Comparison of DynamicVerse with large-scale 2D video datasets and existing 4D scene datasets. DynamicVerse expands the data scale and annotation richness compared to prior works.

Dataset Name	Numerical Statistics			Provided Annotations								Detailed Features			
	#Videos	# Frames	# Masklets	Camera	Depthmap	Instance Mask	Semantic Mask	Object Category	Object Caption	Scene Caption	Camera Caption	Scene Type	Dynamic Type	Real-world?	Metric-scale?
2D Video Dataset															
DAVIS2017	0.2K	10.7K	0.4K	✗	✗	✓	✓	✓	✗	✗	✗	-	-	-	-
Youtube-VIS	3.8K	-	8,171	✗	✗	✓	✓	✓	✗	✗	✗	-	-	-	-
UVO-dense	1.0K	68.3K	10.2K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
VOST	0.7K	75.5K	1.5K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
BURST	2.9K	195.7K	16.1K	✗	✗	✓	✓	✓	✗	✗	✗	-	-	-	-
MOSE	2.1K	638.8K	5.2K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
SA-V	50.9K	4.2M	642.6K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
MiraDATA	330K	-	-	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
4D Scene Dataset															
T.Air Shibuya	7	0.7K	-	✓	✓	✗	✓	✗	✗	✗	✗	Mixed	Street	Synthetic	Yes
MPI Sintel	14	0.7K	-	✓	✗	✓	✗	✓	✗	✗	✗	-	Scripted	Synthetic	-
FlyingThings3D	220	2K	-	✓	✓	✗	✓	✗	✗	✗	✗	Mixed	Objects	Synthetic	-
Waymo	1,150	200K	-	✓	✓	✗	✗	✗	✗	✗	✗	Outdoor	Driving	Real-world	Yes
CoP3D	4,200	600K	-	✓	✗	✓	✗	✗	✗	✗	✗	Mixed	Pets	Real-world	-
Stereo4D	110,000	10,000K	-	✓	✓	✗	✗	✗	✗	✗	✗	Mixed	S. fisheye	Real-world	Yes
PointOdyssey	159	200K	-	✓	✓	✗	✗	✗	✗	✗	✗	Mixed	Realistic	Synthetic	Yes
Spring	47	6K	-	✓	✓	✓	✗	✗	✗	✗	✗	Mixed	Realistic	Synthetic	Yes
Dynamic Replica	524	145K	-	✓	✓	✗	✗	✗	✗	✗	✗	Indoor	Realistic	Synthetic	Yes
MVS-Synth	120	12K	-	✓	✓	✗	✗	✗	✗	✗	✗	Outdoor	Urban	Synthetic	Yes
RealCam-Vid	100K	-	-	✓	✗	✓	✗	✗	✗	✓	✗	Mixed	Realistic	Synthetic	Yes
DynPose-100K	100K	6,806K	-	✓	✗	✗	✗	✗	✗	✗	✗	Mixed	Realistic	Synthetic	Yes
DynamicVerse	100K+	13.6M	800K+	✓	✓	✓	✓	✓	✓	✓	✓	Mixed	Realistic	Real-world	Yes

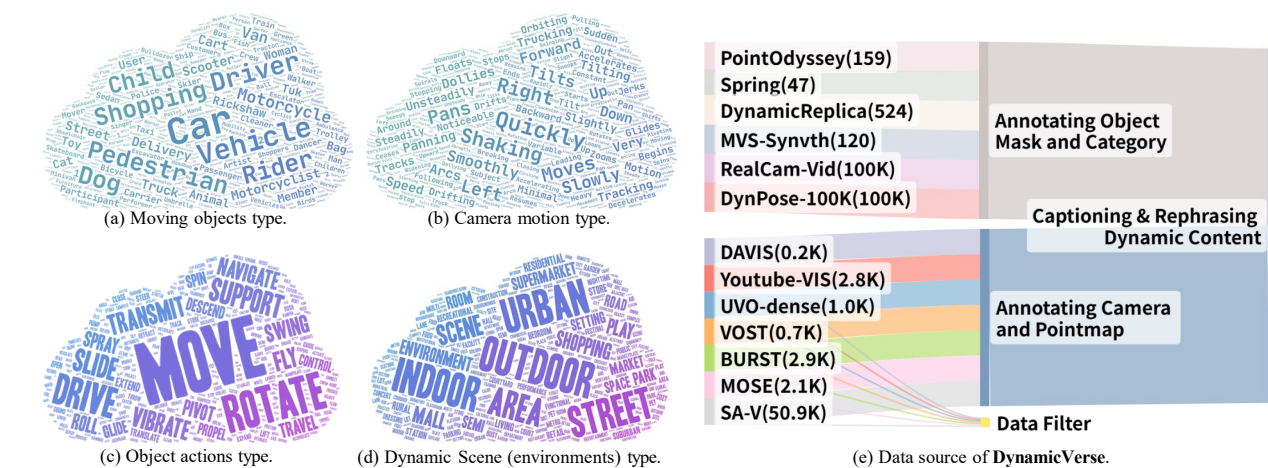


Figure 5. The statistics and data source of DynamicVerse.

