MM3DGS: Multi-modal 3D Gaussian Splatting for SLAM

Lisong C. Sun, Neel P. Bhatt, Jonathan C. Liu, Zhiwen Fan,
Zhangyang Wang, Todd E. Humphreys, Ufuk Topcu

THE UNIVERSITY OF TEXAS AT AUSTIN RADIONAVIGATION LABORATORY http://radionavlab.ae.utexas.edu
CENTER FOR autonomy
TEXAS The University of Texas at Austin

Motivation

From AR/VR to autonomous mobile robotics, Simultaneous Localization and Mapping (SLAM) is an essential enabler and is increasingly being adopted to create 3D scene reconstructions of the operating environment without prior localization.

Point cloud SLAM yields exceptional tracking accuracy, but creates disjoint maps that are visually inferior to newer reconstruction methods. On the other hand, neural radiance field SLAM outputs photorealistic maps, but are computationally expensive and are not scalable due to their implicit nature. 3D Gaussian Splatting (3DGS) addresses these shortcomings with a map representation capable of photorealistic reconstruction and real-time rendering of scenes using multiple posed cameras [1].

This work proposes MM3DGS, a multi-modal SLAM framework that achieves real-time rendering, scale awareness, and improved trajectory tracking with sensor fusion. In addition, a new multi-modal SLAM dataset, UT-MM, is collected from a mobile robot and is publicly released. Experimental evaluation on several scenes from the dataset shows that MM3DGS achieves 3× improvement in tracking and 5% improvement in photometric rendering quality compared to the current 3DGS SLAM state-of-the-art, while allowing real-time rendering of a high-resolution dense 3D map.

[1] Bernhard Kerbl et al. “3d gaussian splatting for real-time radiance field rendering”. In: ACM Transactions on Graphics (ToG) 42.4 (2023), pp. 1–14.

MM3DGS Framework Overview

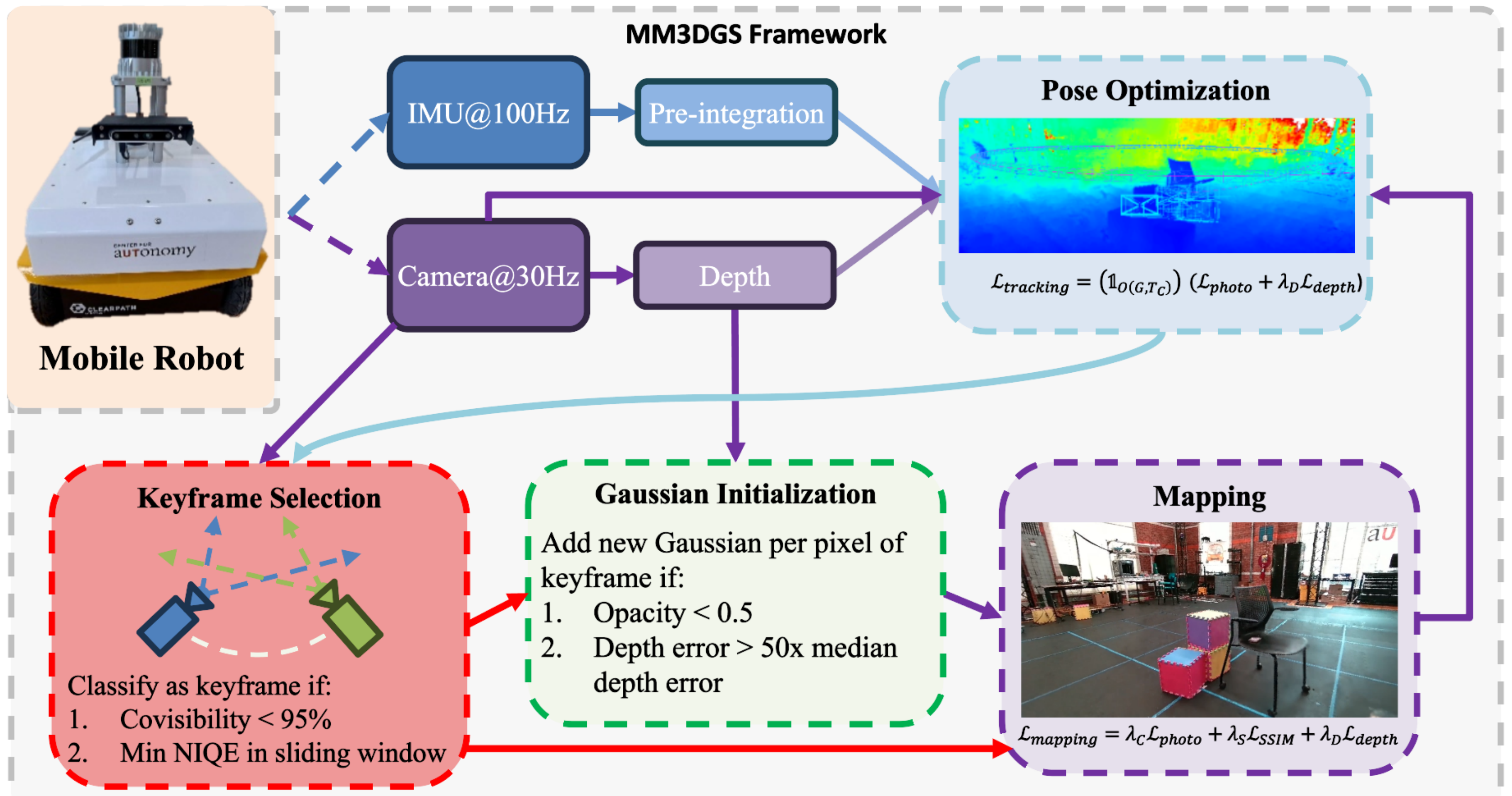

MM3DGS Framework
Mobile Robot
IMU@100Hz
Pre-integration
Camera@30Hz
Depth
Pose Optimization
$\mathcal{L}_{tracking} = (\mathbb{1}_{O(G,T_C)})\,(\mathcal{L}_{photo} + \lambda_D \mathcal{L}_{depth})$
Keyframe Selection
Classify as keyframe if:
1. Covisibility < 95%
2. Min NIQE in sliding window
Gaussian Initialization
Add new Gaussian per pixel of keyframe if:
1. Opacity < 0.5
2. Depth error > 50x median depth error
Mapping
$\mathcal{L}_{mapping} = \lambda_C \mathcal{L}_{photo} + \lambda_S \mathcal{L}_{SSIM} + \lambda_D \mathcal{L}_{depth}$

Figure 1: Overview of the MM3DGS framework. We receive camera images and inertial measurements from a mobile robot. We utilize depth measurements and IMU pre-integration for pose optimization using a combined tracking loss. We apply a keyframe selection approach based on image covisibility and the NIQE metric across a sliding window and initialize new 3D Gaussians for keyframes with low opacity and high depth error [2]. Finally, we optimize parameters of the 3D Gaussians according to the mapping loss for the selected keyframes.

[2] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” IEEE Signal Processing Letters, vol. 20, no. 3, pp. 209–212, 2013.

Methodology

VISUAL SLAM

MM3DGS represents its map as a set of 3D Gaussians $\mathcal{G}$. Each Gaussian has position, covariance, color, and alpha parameters. With an input camera pose T_C, a 2D image can be rendered by splatting the Gaussians on the image plane. Thus, both the Gaussian map and camera poses can be optimized by comparing the L_1 loss between input image I and the render:

$$\mathcal{L}_{\text{photo}} = L_1(I, \text{render}(\mathcal{G}, T_c))$$

Since the map is not guaranteed to cover the entire extent of the current frame, an indicator function is used to control which pixels are optimized:

$$1_{O(\mathcal{G},T_c)} = \begin{cases} 1 & \text{if } O(\mathcal{G},T_c) > 0.99 \\ 0 & \text{otherwise} \end{cases}$$

An additional SSIM loss $\mathcal{L}_{\text{SSIM}}$ aids with mapping.

DEPTH SUPERVISION

Depth measurements can aid SLAM by providing geometric information. The Pearson correlation coefficient is optimized to provide geometric correlation between the measured depth D and rendered depth D_r:

$$\mathcal{L}_{\text{depth}} = \frac{\text{Cov}(D, D_r)}{\sqrt{\text{Var}(D)\text{Var}(D_r)}}$$

Further, new Gaussians can be initialized at the sensed depth to cover unseen areas.

INERTIAL FUSION

Prior to pose optimization, a good initial estimate is necessary for preventing divergence to bad local minima. IMU measurements can be integrated to accurately propagate the camera pose between frames. While this open-loop chaining does not account for IMU biases, we find that errors are small enough to be optimized away by the visual SLAM.

UT-MM Dataset

To test the MM3DGS framework, a multi-modal dataset dubbed UT Multi-modal (UT-MM) was collected. The dataset includes eight scenes with RGB-D images at 30 Hz, inertial measurements at 100 Hz, and LiDAR point clouds at 10 Hz.

UT-MM addresses the lack of visual-inertial SLAM datasets with high quality RGB images. We aim for UT-MM to be a benchmark for future multi-modal photorealistic SLAM frameworks.

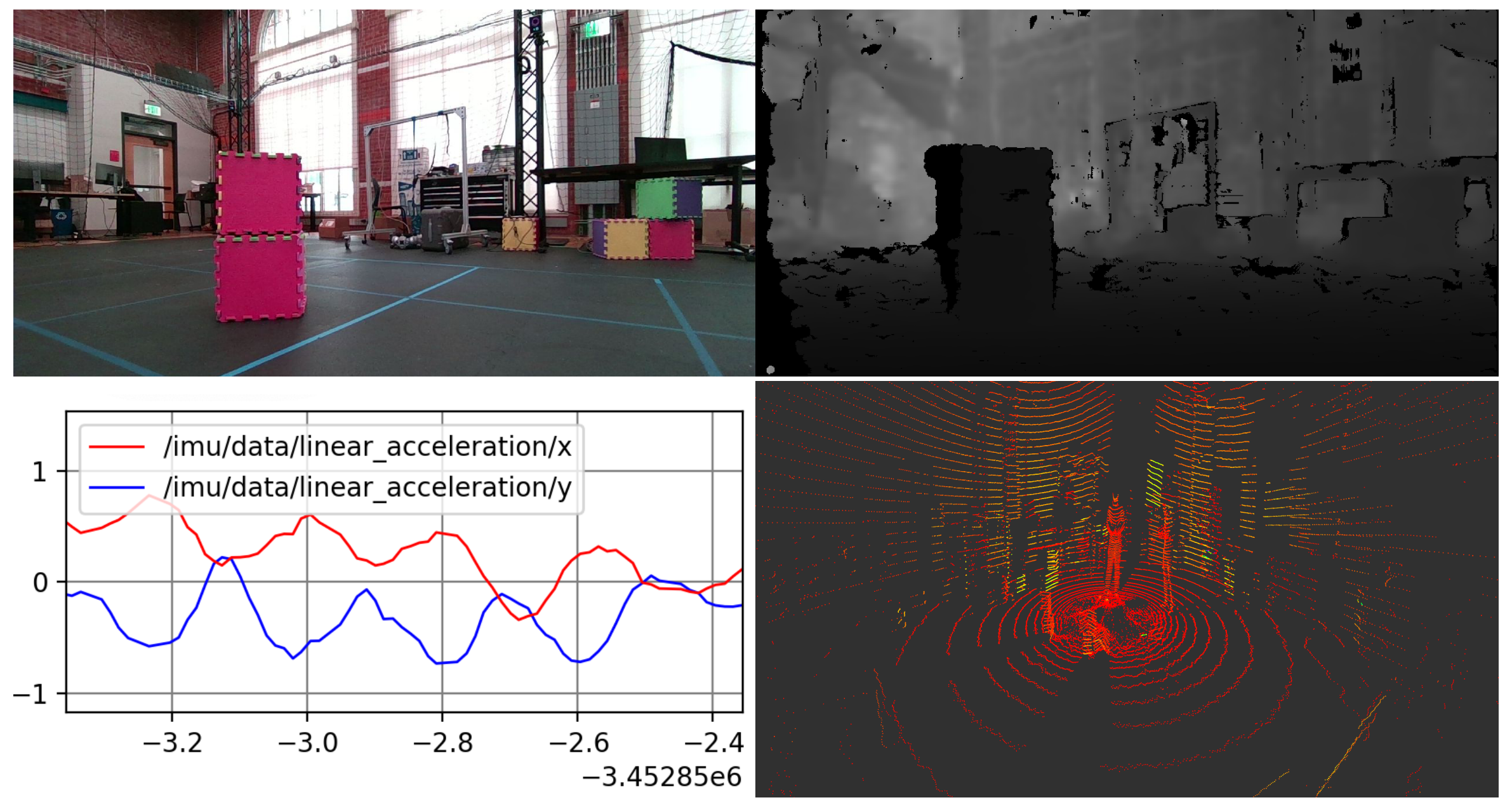

/imu/data/linear_acceleration/x
/imu/data/linear_acceleration/y
1
0
−1
−3.2
−3.0
−2.8
−2.6
−2.4
−3.45285e6

Figure 2: Our dataset provides RGB images (top left), depth images (top right), IMU measurements (bottom left), and LiDAR point clouds (bottom right). The above examples are taken from the Ego-drive scene.

Results

Method	Avg		Square-1		Ego-centric-1		Ego-drive		Fast-straight	
	ATE	PSNR	ATE	PSNR	ATE	PSNR	ATE	PSNR	ATE	PSNR
SplaTAM (RGB-D)	12.06	22.03	32.86	**18.67**	4.40	22.78	**4.20**	20.61	6.78	26.07
Ours (RGB)	39.14	19.73	59.48	16.54	4.09	23.151	67.20	17.51	25.78	21.71
Ours (RGB+IMU)	33.23	19.58	44.26	17.01	3.41	22.96	68.50	17.12	16.78	21.24
Ours (RGB-D+IMU)	**3.98**	**23.30**	**7.11**	18.59	**1.15**	**24.95**	4.54	**23.61**	**3.13**	26.05

Table 1: Multi-modal SLAM results on the UT-MM dataset: ATE RMSE ↓ is in cm and PSNR ↑ is in dB, with SplaTAM is used as a baseline. Best results are in **bold**. Both depth and inertial measurements benefit tracking and image quality.

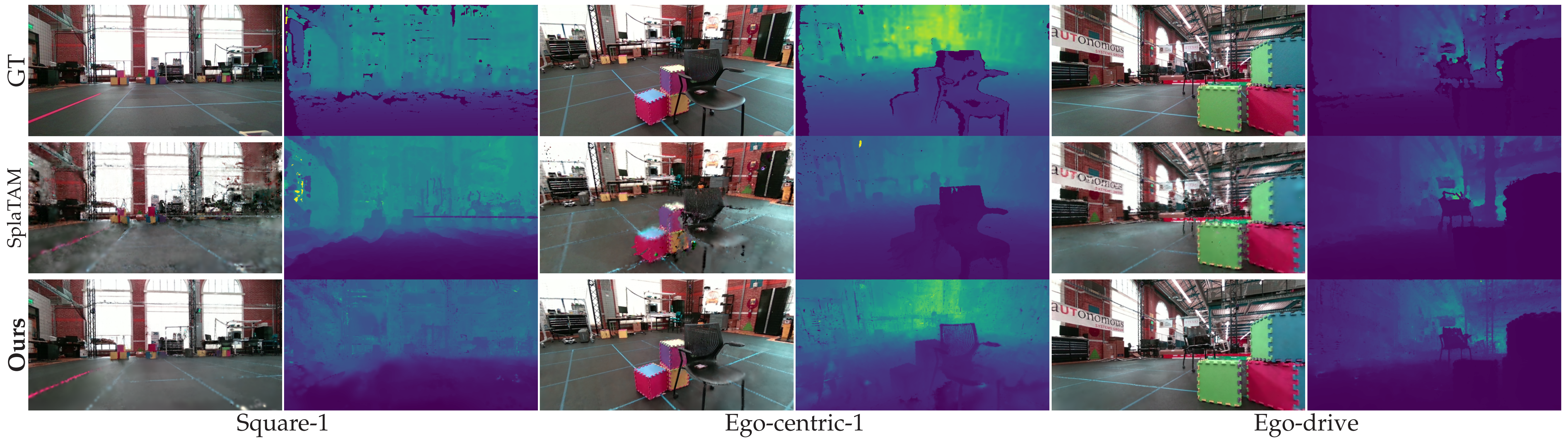
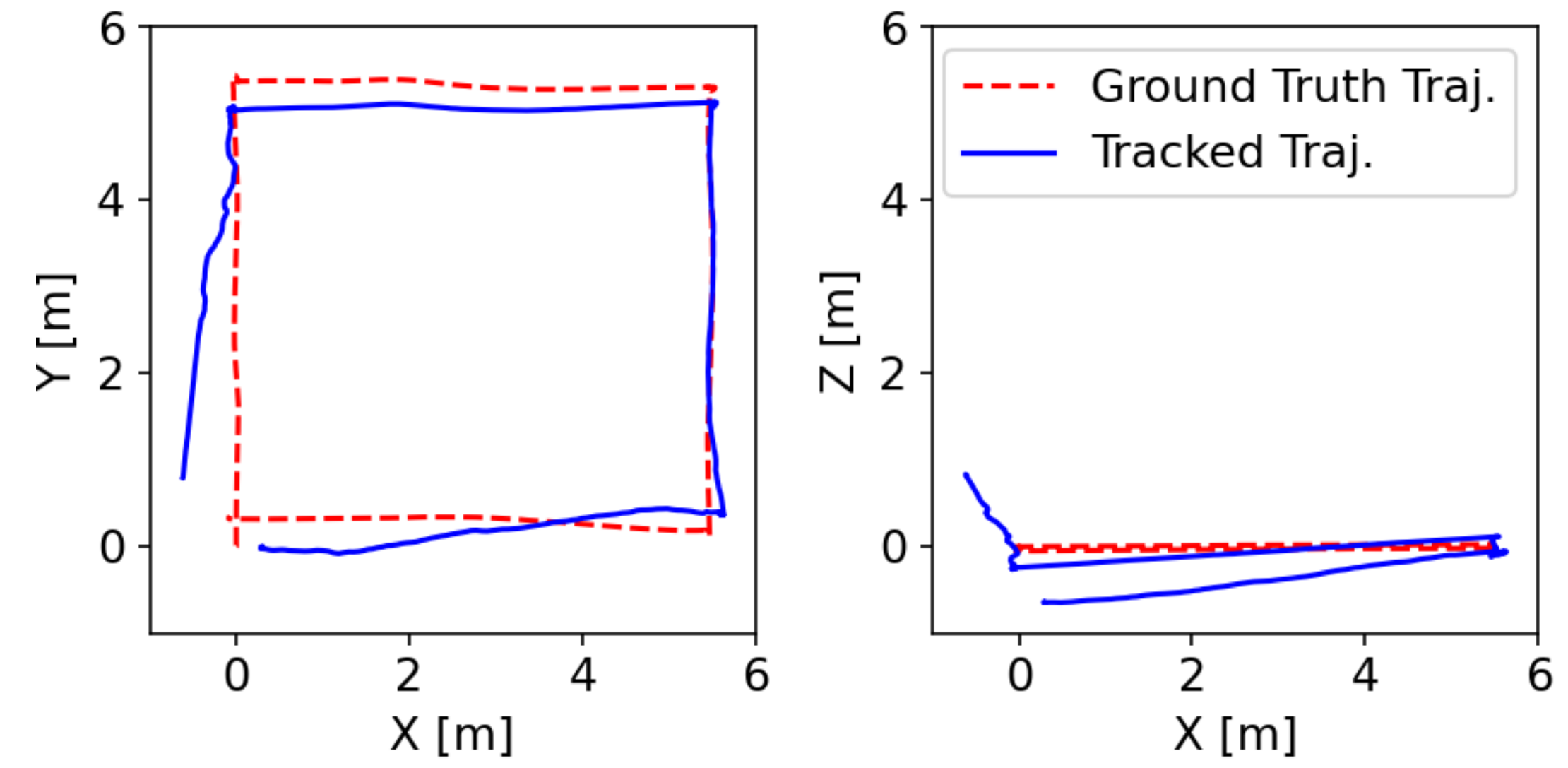


Figure 3: Qualitative results on UT-MM dataset: RGB and depth renderings of UT-MM scenes. Note that the ground truth (GT) depths are captured with depth cameras, and thus are imperfect. Our method exhibits geometric details not present in the GT depth, as well as fewer RGB artifacts compared to SplaTAM.

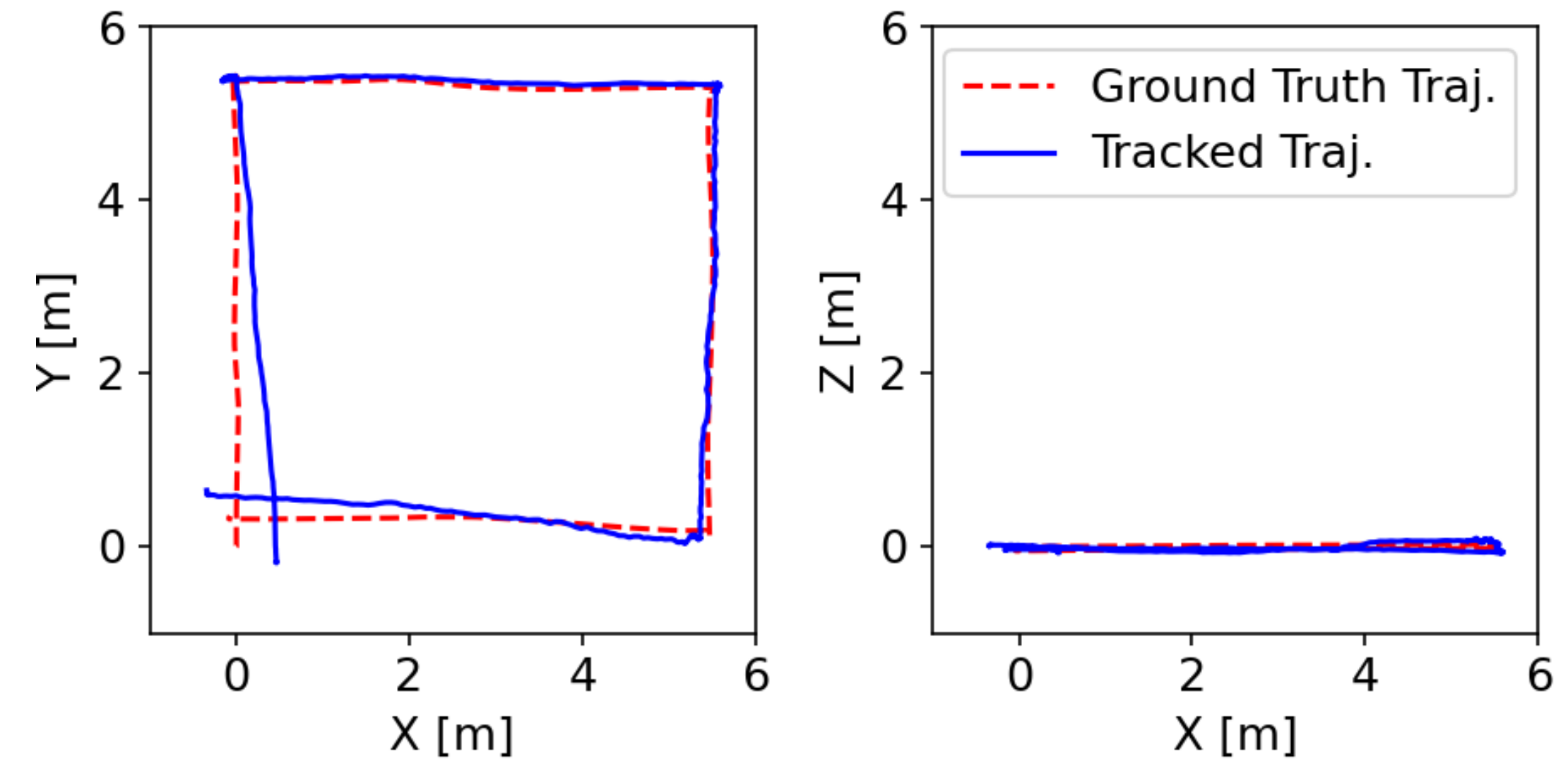
Effects of Multi-modal Tracking

To examine the effects of sensor fusion, the UT-MM Square-1 scene is tracked with various sensor configurations.

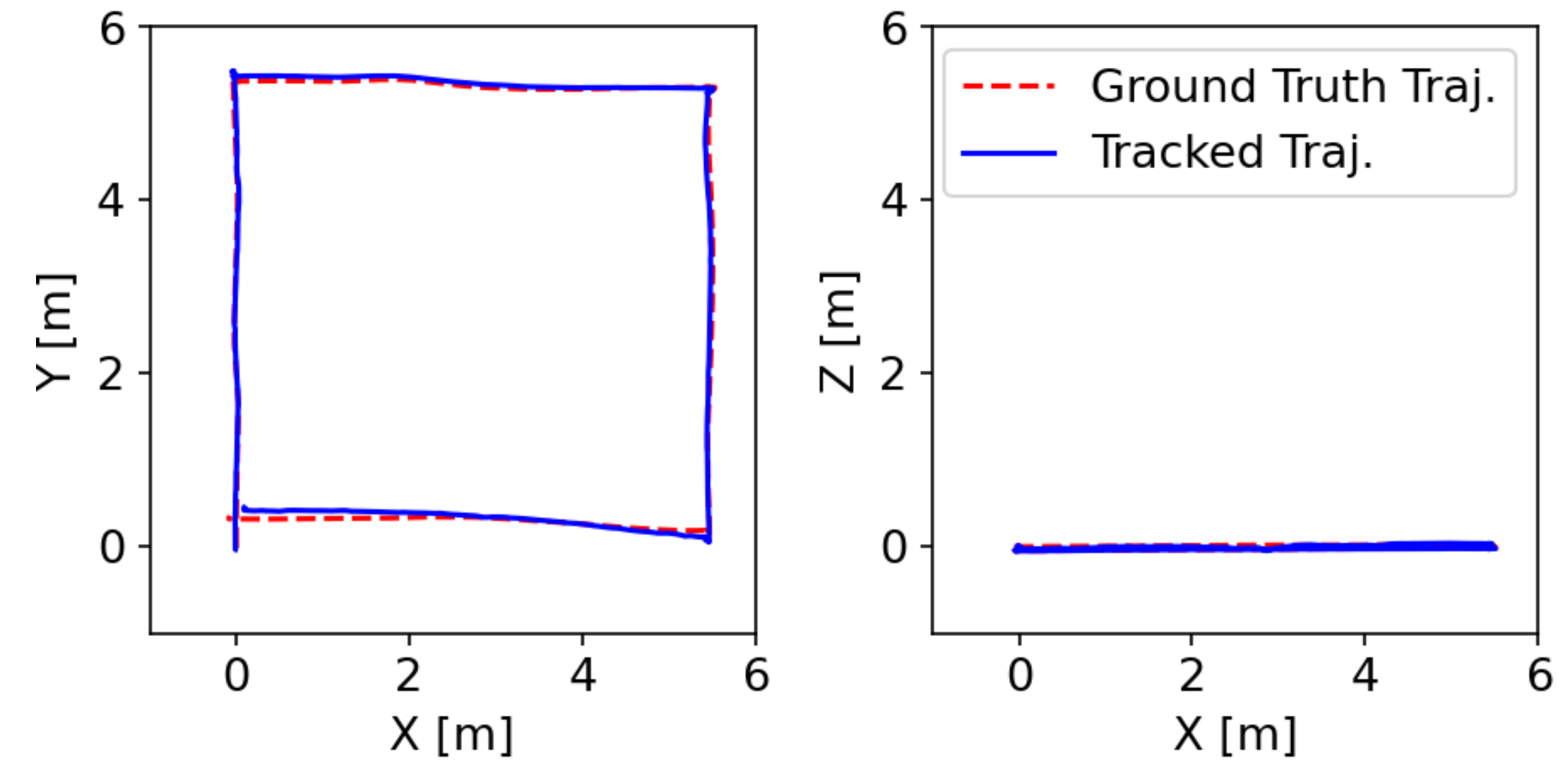
RGB ONLY



RGB + DEPTH



RGB + DEPTH + IMU



Next Steps

Future work aims to improve MM3DGS by adding

1. Robustness with tightly-coupled inertial fusion and bias estimation
2. Scalability with loop closure
3. Speed with advanced 3DGS techniques, e.g., DUSt3R and InstantSplat

Conclusion

We presented MM3DGS, a multi-modal SLAM framework built on a 3D Gaussian map representation. We evaluate our framework on a new multi-modal dataset, UT-MM, that includes RGB-D images, IMU measurements, LiDAR depth, and ground truth trajectories. MM3DGS achieves superior tracking accuracy and rendering quality compared to the state-of-the-art baseline. MM3DGS can be implemented in a wide range of applications in robotics, augmented reality, and mobile computing due to its use of commonly available sensors.

See our project page!