WORLD LAUREATES FORUM

世界顶尖科学家论坛

国际工程智能大会

2026 International Conference on AI for Engineering

World Laureates Foundation

上海世界顶尖科学家发展基金会

2026

工程智能研究院

INSTITUTE OF AI FOR ENGINEERING

VGGT-Motion: Motion-Aware Calibration-Free Monocular SLAM for Long-Range Consistency

ICML International Conference On Machine Learning

Zhuang Xiong[1*], Chen Zhang[1*], Qingshan Xu[2], Wenbing Tao[1]

[1] Huazhong University of Science and Technology, Wuhan [2] University of Science and Technology of China, Hefei

01 ABSTRACT

VGGT-Motion pairs motion-aware submaps with context-balanced anchors to keep calibration-free monocular SLAM globally consistent over long sequences.

Research question

How can monocular SLAM stay globally consistent over kilometer-scale trajectories without known camera calibration?

02 METHOD

Optical flow identifies static and turning segments. Parallax-based sampling removes redundancy; context-balanced anchors directly register dense submaps in Sim(3).

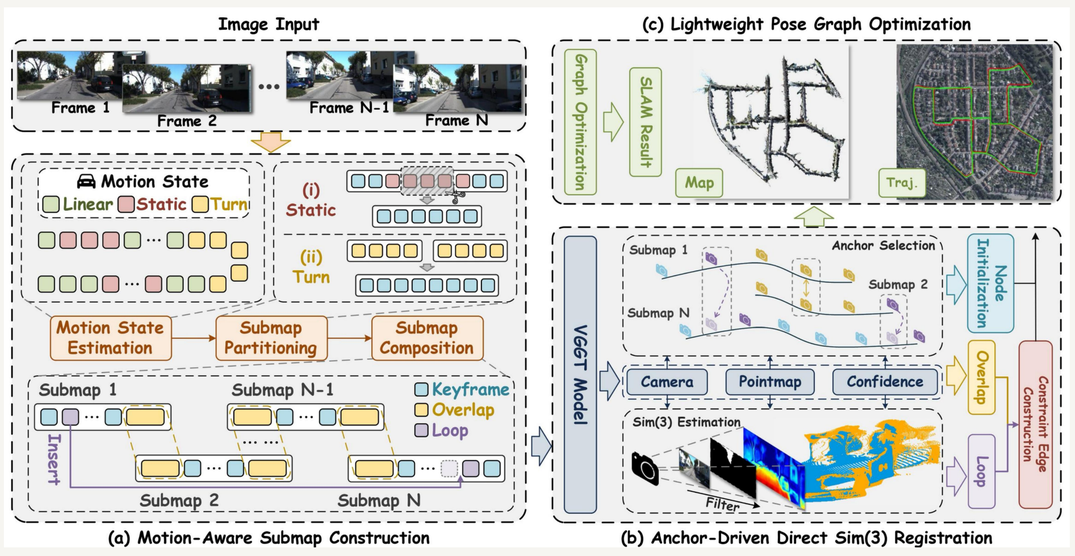


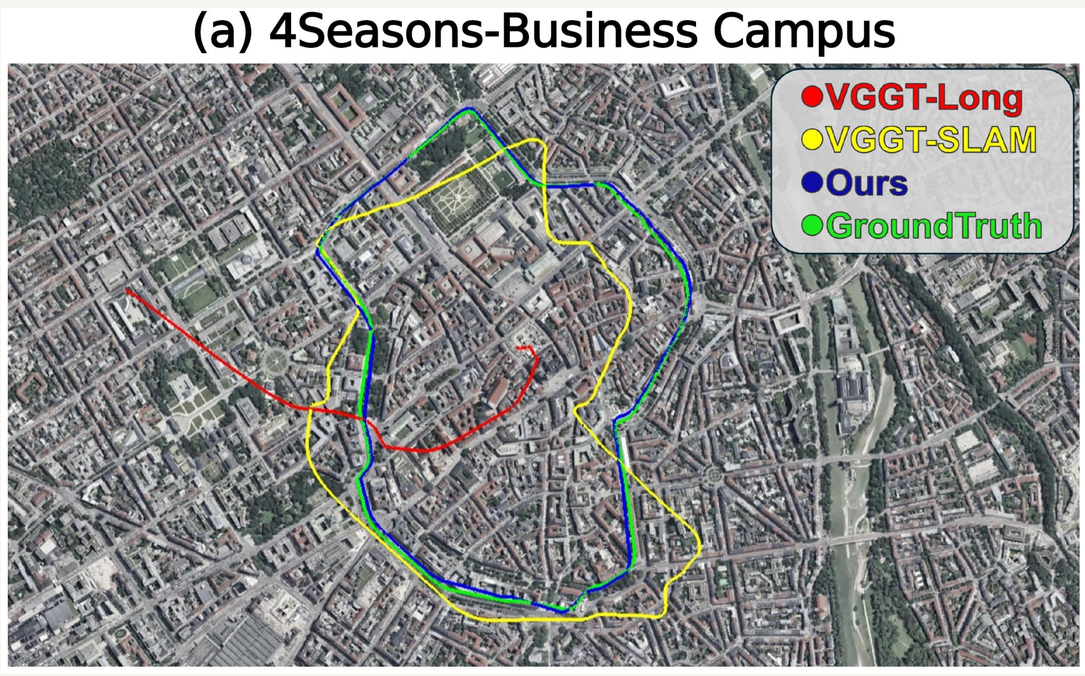
Figure 1. Motion-aware submaps, direct Sim(3) registration and lightweight pose graph optimization.

Evaluation protocol

Five outdoor datasets; ATE RMSE measured after global Sim(3) alignment.

03 RESULTS

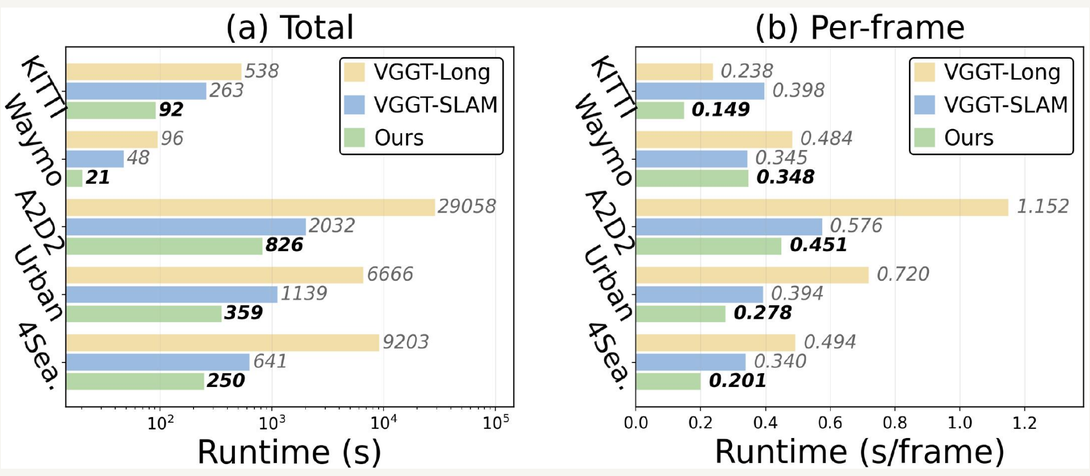
On three unseen long-range datasets, VGGT-Motion reduces ATE by **84–96%** versus VGGT-Long and runs **18–36×** faster.



Paper Fig. 4a. 4Seasons trajectory.

Method	*Avg Frame: 16,495* 4Seasons ATE (m) ↓	4Seasons Drift (%) ↓	*Avg Frame: 12,931* Complex Urban ATE (m) ↓	Complex Urban Drift (%) ↓	*Avg Frame: 24,842* A2D2 ATE (m) ↓	A2D2 Drift (%) ↓
CUT3R	OOM	OOM	OOM	OOM	OOM	OOM
Fast3R	OOM	OOM	OOM	OOM	OOM	OOM
MASt3R-SLAM	TL	TL	TL	TL	TL	TL
VGGT-SLAM	168.04	4.72%	444.63	7.73%	192.80	6.00%
VGGT-Long	280.25	7.16%	475.59	8.20%	182.93	5.69%
Ours	**12.22**	**0.32%**	**35.48**	**0.58%**	**29.80**	**0.93%**

Paper Table 3. ATE/drift; OOM and TL mark failed runs.



Paper Fig. 5. Total and per-frame runtime.

04 DISCUSSION

Offline runtime still depends on VGGT inference. Highly dynamic scenes and long intervals without loop closure remain challenging.

05 CONCLUSIONS

Motion-aware input selection and context-balanced anchors support accurate, efficient long-range monocular SLAM without known intrinsics.

CONTACT

qingshan.xu@ustc.edu.cn · wenbingtao@hust.edu.cn

REFERENCES / ACKNOWLEDGEMENTS

Xiong et al., VGGT-Motion, ICML 2026, PMLR 306.