

## A More Experimental Results

**Time Evaluation.** The entire model training process lasted approximately 4 hours, with a peak GPU memory consumption of around 10 GB. During inference, the model demonstrates efficient real-time processing capabilities. On the Oxford/QEOxford datasets, the average processing time per sample is 29ms (34 FPS), and on the NCLT dataset, it is 48ms (21 FPS). These speeds are well within the respective scanning frequencies of 20 Hz for Oxford/QEOxford and 10 Hz for NCLT, highlighting the model’s ability to maintain high accuracy during real-time operation.

**More Dataset Details.** Regarding the datasets, ground truth poses for the Oxford [1] and QEOxford [3] datasets are obtained through interpolation from an integrated GPS/INS system. For the NCLT [4] dataset, ground truth poses are generated post-collection using SLAM. Further details on the data splits can be found in Tab. 1 and Tab. 2.

**More Ablation Study.** To evaluate our system’s ability to generate text for inference without requiring ground-truth poses, we conduct an ablation study that incorporates an auxiliary classification network specifically for this purpose. This network is designed with a LightLoc [2] encoder, followed by four fully-connected layers of identical feature dimensions, to separately predict partitions for position and orientation. The training configurations are consistent with those used for GTR-Loc. During inference, this classification network first predicts districts  $z$  and directions  $d$ . This prediction is then used to generate a conditioned textual input for our primary localization network, depicted as ‘with text (cls)’. The detailed localization results alongside the position/orientation classification accuracy of this auxiliary network are presented in Tab. 3 and Tab. 4. The results indicate that this approach delivers virtually no performance gains, while it increases both test time and computational complexity owing to the additional classification network and the LiDAR-text regression. The limited classification accuracy on the specific dataset, i.e., NCLT, produces faulty text outputs, which in turn propagate errors into the localization results.

Table 1: Details of the Oxford dataset.

Sequence	Length (km)	Weather	Split
11-14-02-26	9.37	sunny	Train
14-12-05-52	9.22	overcast	Train
14-14-48-55	9.05	overcast	Train
18-15-20-12	9.04	overcast	Train
15-13-06-37	8.85	overcast	Eval
17-13-26-39	9.02	sunny	Eval
17-14-03-00	9.02	sunny	Eval
18-14-14-42	9.04	overcast	Eval

Table 2: Details of the NCLT dataset.

Sequence	Length (km)	Weather	Split
2012-01-22	6.10	overcast	Train
2012-02-02	6.20	sunny	Train
2012-02-18	6.20	sunny	Train
2012-05-11	6.00	sunny	Train
2012-02-12	5.80	sunny	Eval
2012-02-19	6.20	overcast	Eval
2012-03-31	6.00	overcast	Eval
2012-05-26	6.30	sunny	Eval

Table 3: Ablation of using text at inference.

Method	QEOxford	Oxford	NCLT
vanilla	0.83/1.12	2.67/1.25	1.46/2.80
<b>ours</b>	0.75/1.03	2.59/1.19	1.40/2.62
with text (cls)	0.75/1.02	2.60/1.21	1.41/2.70

Table 4: Classification accuracy of position/orientation on different datasets.

Dataset	Accuracy
QEOxford	99.19%/97.64%
Oxford	98.99%/97.52%
NCLT	98.44%/85.50%

**More visualization.** To further dissect the performance on the Oxford and NCLT datasets, Fig. 1 and Fig. 2 provide trajectory visualizations and cumulative error distribution curves. The results on sequences 17-13-26-39 (Oxford) and 2012-02-19 (NCLT) are provided for comparison, respectively. GTR-Loc’s estimated trajectory adheres closely to the ground truth throughout different sequences. Even within structurally complex or repetitive areas (as marked by blue boxes), GTR-Loc maintains consistent localization. Methods like SGLoc or LiSA exhibit jumps in these areas. The cumulative error distribution curves also demonstrate GTR-Loc’s leading performance across most error ranges. Our curves for both position and orientation errors predominantly lie above those of other methods, suggesting overall lower error magnitudes.

## 34 B Limitations and Future Work

35 **Limitation.** Despite its promising results, GTR-Loc has limitations that highlight areas for future  
 36 research. While our distillation approach effectively eliminates the need for text processing during  
 37 inference, a limitation is that leveraging text directly at inference time demonstrably achieves better  
 38 performance. Our ablation studies in the main paper underscore this point. In contrast, the appendix  
 39 shows that the straightforward strategy of using a classification network to predict position and  
 40 orientation categories for geospatial text generation is inadequate.

41 **Future Work.** Consequently, our future work will concentrate on exploring methods for generating  
 42 more accurate textual descriptions without using ground truth poses during the inference phase. The  
 43 aim is to significantly enhance LiDAR-text localization precision by effectively harnessing these  
 44 improved textual cues at the point of decision-making.

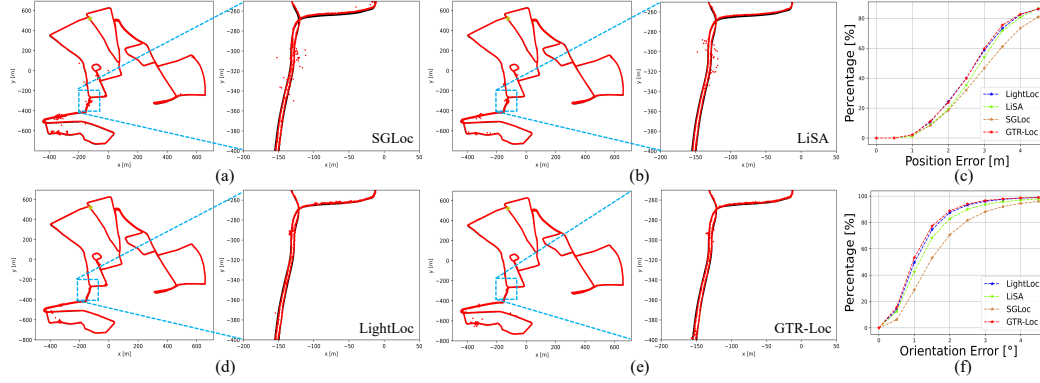


Figure 1: Visual comparisons on Oxford. (a) (b) (d) (e): predicted trajectories (red) overlaid on ground truth (black); a star marks the starting position, and the blue box highlights a challenging road segment. (c) (f): cumulative error distribution curves for position (top) and orientation (bottom).

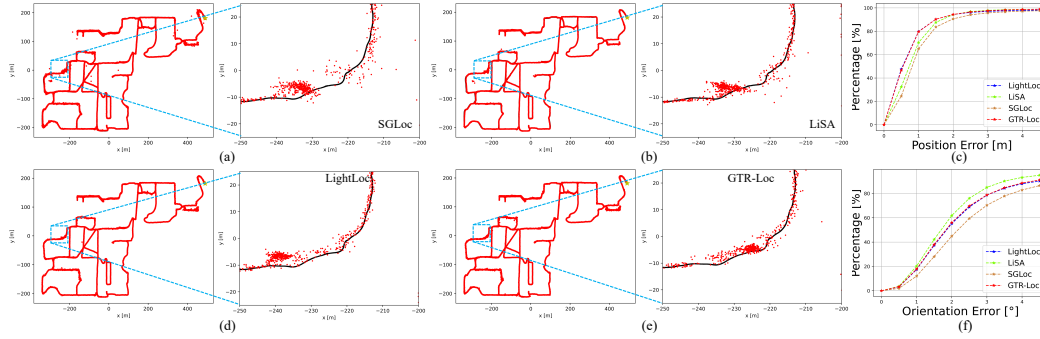


Figure 2: Visual comparisons on NCLT. The notations follow Fig. 1.

## 45 References

- 46 [1] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar  
 47 robotcar dataset: A radar extension to the oxford robotcar dataset. In *ICRA*, pages 6433–6438,  
 48 2020.
- 49 [2] Wen Li, Chen Liu, Shangshu Yu, Dunqiang Liu, Yin Zhou, Siqi Shen, Chenglu Wen, and Cheng  
 50 Wang. Lightloc: Learning outdoor lidar localization at light speed. In *CVPR*, 2025.
- 51 [3] Wen Li, Shangshu Yu, Cheng Wang, Guosheng Hu, Siqi Shen, and Chenglu Wen. Sgloc: Scene  
 52 geometry encoding for outdoor lidar localization. In *CVPR*, pages 9286–9295, 2023.
- 53 [4] Carlevaris-Bianco Nicholas, K. Ushani Arash, and M. Eustice Ryan. University of michigan  
 54 north campus long-term vision and lidar dataset. *IJRR*, 35:545–565, 2015.