Egocentric Action-aware Inertial Localization in Point Clouds with Vision-Language Guidance

Supplementary Material

9. Supplementary Ablation Studies

Different Vision-Language Encoder in Stage 1 We compare CLIP-Large[50], InternVL[7], and SLIP-Base[43] as possible backbones for the vision-language encoders. While InternVL scores the highest on some metrics, SLIP generally attains stronger localization accuracy. Hence, in our default configuration, we adopt SLIP as the vision-language encoder.

Temporal Length of Input IMU Sequence in Stage 2 We experiment with sequences of 5s, 10s, and 20s of IMU data. Shorter 5s windows excel on the seen rooms but underperform on the unseen set. Longer 20s windows better preserve temporal context and yield a slight gain in unseen action top-1 performance, yet they degrade localization in both settings. By contrast, 10s serves as a balanced choice, delivering robust localization and action recognition across seen/unseen scenarios.

Residual Connection in Stage 2 Architecture Design Removing residual connections, either on the IMU feature path or on the point cloud feature path, consistently degrades results. In contrast, preserving both forms of residual connections significantly improves localization and action accuracy. This highlights the importance of allowing the network to fuse new spatiotemporal cues while maintaining a direct pathway for unimodal features.

Preliminary Location Retrieval Accuracy in Stage 1 Lastly, we assess how well the short-term feature alignment from Stage 1 alone can recover the user's location. Doing so attains considerably lower accuracy compared to our full Stage 2 inference. This underscores the utility of leveraging both temporal context and global spatial reasoning to refine the initial retrieval from Stage 1.

10. Supplementary Video

We create a supplementary video that further illustrates the performance of our framework. The heatmaps featured in the video are interpreted similarly to those shown in Fig. 5, providing visual insight into the localization process. The video demonstrates the robustness of our approach, highlighting its capability for accurate long-term human tracking across extended periods.

Table 5. Supplementary Ablation Studies.

Method	Seen Rooms						Unseen Rooms					
	0.2m	0.4m	0.6m	RS	$\mathcal{A}$-top1	$\mathcal{A}$-top5	0.2m	0.4m	0.6m	RS	$\mathcal{A}$-top1	$\mathcal{A}$-top5
Different Vision-Language Encoder in Stage 1												
CLIP	42.72	69.33	89.63	95.94	20.81	52.05	21.60	53.45	83.41	86.44	12.68	40.38
InternVL	42.41	69.21	**90.95**	**96.15**	17.15	50.80	26.11	60.46	87.70	**90.18**	9.84	39.44
SLIP	**43.86**	**70.15**	89.6	96.01	**21.48**	**53.62**	**26.86**	**65.97**	**90.79**	89.55	**15.03**	**43.34**
Temporal Length of Input IMU Sequence in Stage 2												
5s	**45.27**	**71.86**	**90.84**	**96.18**	21.23	53.57	25.69	64.07	87.11	89.28	14.53	43.62
20s	43.10	69.61	89.11	95.84	20.82	52.53	24.77	61.09	85.99	89.15	**15.41**	**44.77**
10s	43.86	70.15	89.6	96.01	**21.48**	**53.62**	**26.86**	**65.97**	**90.79**	**89.55**	15.03	43.34
Residual Connection in Stage 2 Architecture Design												
w/o IMU residual	36.29	63.56	86.84	95.05	21.45	52.38	26.07	60.41	86.03	89.10	14.90	43.15
w/o PC residual	39.14	64.71	86.16	95.30	20.24	52.08	23.51	59.66	86.02	87.78	14.60	43.77
w/ both	**43.86**	**70.15**	**89.6**	**96.01**	**21.48**	**53.62**	**26.86**	**65.97**	**90.79**	**89.55**	**15.03**	**43.34**
Preliminary Location Retrieval Accuracy in Stage 1												
Stage 1 Retrieval	18.29	41.23	66.70	85.47	/	/	14.25	38.51	64.31	78.31	/	/
Full 2 Stages	**43.86**	**70.15**	**89.6**	**96.01**	**21.48**	**53.62**	**26.86**	**65.97**	**90.79**	**89.55**	**15.03**	**43.34**