MVP-SLAM: Multi-View Visual-Inertial Plan-Prior SLAM

Asier Bikandi-Noya[1,*], Miguel Fernandez-Cortizas[1,*], Muhammad Shaheer[1], Holger Voos[1], and Jose Luis Sanchez-Lopez[1]

***Abstract*—This report describes our submission to the Hilti × Trimble SLAM Challenge 2026 [1]. We present MVP-SLAM, a visual-inertial SLAM system for the Insta360 One-RS dual-fisheye sensor that uses the as-planned floorplan of the construction site as a prior. The system extends ORB-SLAM3 with a multi-camera tracking front-end and integrates floorplan walls into the optimization through a global pose-graph that anchors the trajectory to the building's as-planned design.**

I. System Overview

Our system builds on ORB-SLAM3 [2] and adds two contributions: a multi-camera visual-inertial front-end for the dual-fisheye sensor, and a floorplan integration module that turns the as-planned drawing into geometric constraints on the trajectory. The latter extends our earlier work on BIM-informed SLAM for construction monitoring [3] from RGB-D to a visual-inertial dual-fisheye setup.

A. Multi-camera Visual-Inertial SLAM

We extend ORB-SLAM3 to handle the two fisheye cameras of the Insta360 sensor. Each map point is shared between the two cameras and tracked from whichever one observes it, so a point first triangulated by one camera can be re-observed by the other when the sensor rotates. Both cameras contribute to map maintenance, place recognition, and bundle adjustment.

We adapt the initialization to the construction-site capture pattern, which often starts with the sensor stationary or moving slowly. Initialization timeouts and triggers are tuned to fire sooner, and a guard prevents the stationary-start map reset that ORB-SLAM3 would otherwise apply. We also revised the camera–IMU calibration, which the system is sensitive to in the fisheye case.

To produce a trajectory that covers the full sequence, required to satisfy the challenge's 99 % coverage rule, we fill in the pre-initialization frames with stationary (ZUPT) poses anchored to the first keyframe, and we align the trajectory using whichever map in the atlas contains the initial-pose timestamp, so sequences with mid-run tracking restarts still produce a complete trajectory file.

*These authors contributed equally.

[1]Authors are with the Automation and Robotics Research Group, Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg. Holger Voos is also associated with the Faculty of Science, Technology, and Medicine, University of Luxembourg, Luxembourg. {asier.bikandi, muhammad.shaheer, miguel.fernandez, holger.voos, joseluis.sanchezlopez}@uni.lu

TABLE I
APE RMSE (m) AND CHALLENGE SCORE ON THE FIVE EARLY-RELEASE SEQUENCES FOR THE SLAM AND LOCALIZATION TASKS. SCORES ARE FROM THE OFFICIAL EVALUATION SERVER. COVERAGE IS 100 % ON EVERY RUN.

	SLAM		Localization	
Sequence	APE	Score	APE	Score
floor_1_2025-05-05_run_1	0.405	85.2	0.456	83.1
floor_2_2025-05-05_run_1	0.538	81.0	0.315	89.9
floor_2_2025-10-28_run_1	0.187	92.5	0.188	93.1
floor_2_2025-10-28_run_2	0.163	93.6	0.247	90.9
floor_UG1_2025-10-16_run_1	0.562	81.1	0.550	82.8

B. Floorplan Integration

The DXF floorplan of the building is parsed once into a set of wall segments expressed in the floorplan frame. The trajectory is anchored to this frame using a rigid transform computed from the published $t \approx 5\,\mathrm{s}$ initial pose; the transform is recomputed whenever the inertial bundle adjustment rescales the map.

Wall priors are produced from the camera stream by semantic segmentation. We use only the WALL class, which is the most stable label across construction-site lighting and texture. Map points whose projections fall on WALL pixels in either camera are tagged accordingly, and a small reinforcement-and-erosion counter stabilises the label against per-frame segmentation flicker. Wall-tagged map points are passed at keyframe rate to a wall-fitting thread that runs sequential RANSAC, with the IMU's gravity vector used to constrain the plane normals. The result is a set of wall planes recovered online from the sparse map.

Detected walls are matched against floorplan walls by an association step that compares azimuth, distance, centroid separation, and length ratio. Adaptive gates distinguish between interior and exterior walls, with the interior gates being the tighter of the two. The distance gate also adapts to the trajectory length since the last correction, so it is tight at the start of a run and widens after long traversals where larger drift is expected, preventing both false matches early on and missed associations after long stretches without floorplan support. The matched pairs feed a global pose-graph that contains the trajectory keyframes, the detected walls, and the as-planned floorplan walls. The floorplan walls enter as fixed anchors; the detected walls are optimised toward them; and the keyframes move to remain consistent with their wall observations. The

corrected keyframe poses are then injected back as unary factors into the local inertial bundle adjustment, so the rest of the system is pulled toward the floorplan-anchored solution.

C. BEV-based Refinement

The wall integration of Section I-B operates at keyframe rate on sparse map points, leaving short-horizon drift between keyframes uncorrected. We add a complementary per-frame module that densifies the wall signal into a single-image bird's-eye view (BEV) and registers it directly to the floorplan.

For each image column, we extract the floor–wall and the wall–ceiling boundary pixels from the same segmentation network used in Section I-B and lift their elevation angles, expressed in the gravity-aligned frame, into per-column distances to the nearest wall. Because the two boundaries are corrupted by different sources such as floor clutter, lights and ducts etc. we build a BEV from each boundary separately, dry-run ICP against the floorplan on both, and select whichever yields more matches at lower residual. The selected BEV is then registered to the floorplan walls by a 2D ICP in (x, y, yaw), warm-started from the current SLAM pose (see Fig. 2).

Several details stabilize the registration. The floorplan walls are pre-filtered to keep only the innermost face of each thick wall, since the BEV cloud always lies inside the room and the outer face would otherwise attract spurious matches. The ICP gates are adapted to the scene: closed rooms use a loose-then-tight two-pass schedule, while open scenes such as parking garages, where the BEV is dominated by isolated columns rather than continuous walls, use a single tighter pass, with the choice driven by the residual distribution of a loose dry-run. Per-frame corrections that exceed a SLAM-conditioned validation gate are rejected so that ambiguous matches do not propagate; the accepted corrections are inserted as unary factors into the global pose-graph of Section I-B and re-optimised together with the keyframe-level wall constraints.

II. Experimental Results and Discussion

We evaluate on the five sequences for which the organisers released ground truth, using the official scoring formula. The first five seconds of each run are excluded per the challenge protocol. Coverage is 100 % on every sequence.

Table I reports APE RMSE and the score for the two challenge tasks. The SLAM column reports the trajectory in any frame, rigid-aligned to ground truth via Umeyama. The Localization column reports the trajectory directly in the floorplan frame, evaluated with no alignment. Figure 1 shows the system running on `floor_2_2025-05-05_run_1`: the full trajectory, the wall-tagged map points, the detected wall planes, and the as-planned floorplan overlay.

III. Conclusion

We presented MVP-SLAM, our submission to the Hilti × Trimble SLAM Challenge 2026. The system extends ORB-SLAM3 to a dual-fisheye visual-inertial setup and integrates the as-planned floorplan as a geometric prior through wall detection, association, and a global pose-graph that anchors the trajectory to the building's design. The same pipeline produces the trajectory for both the SLAM and Localization tasks.

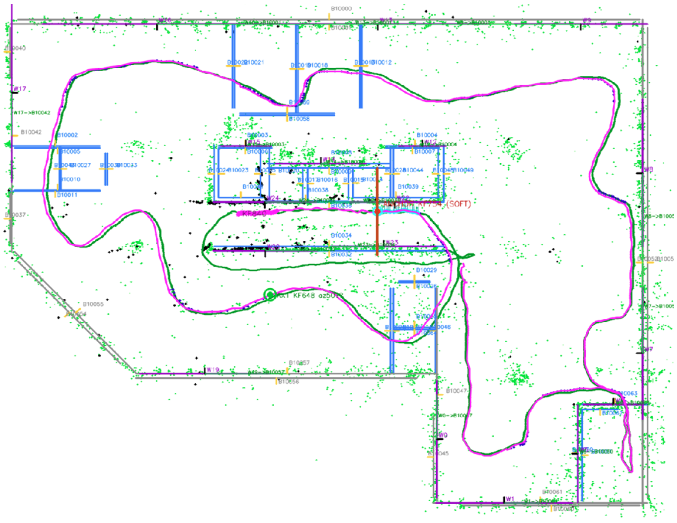

Fig. 1. Visualisation of the system on `floor_2_2025-05-05_run_1`. Estimated trajectory in red, ground truth in green, semantic-coloured map points (green = WALL), detected walls and as-planned floorplan in blue.

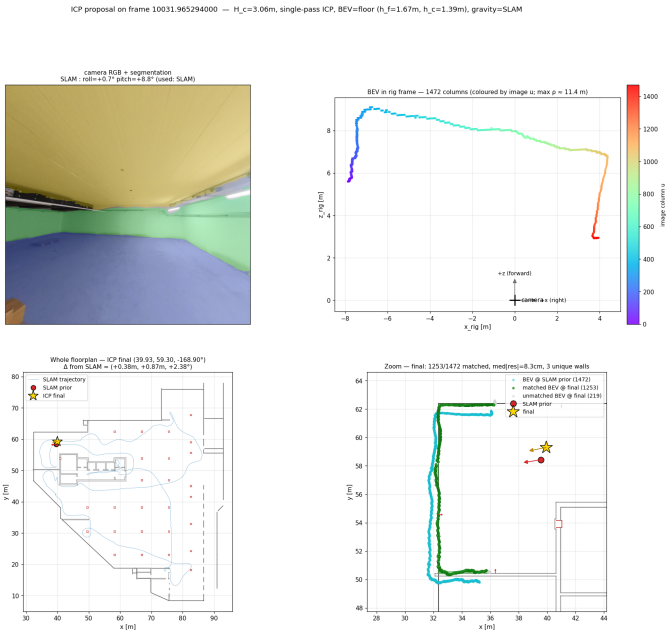

Fig. 2. Per-frame BEV refinement on a representative frame of `floor_UG1_2025-12-02_run_1`. The BEV scan extracted from the floor-wall boundary (top right) is registered to the as-planned floorplan by 2D ICP in (x, y, yaw), warm-started from the SLAM prior. The zoom panel shows 1253 of 1472 BEV columns associating with three floorplan walls at a median residual of 8.3 cm, yielding a 0.95 m translation correction and 2.4° yaw correction over the keyframe-level pose.

References

[1] S. Centanni, Y. Zhang, Y. Tao, J. Kindle, F. Neuhaus, T. Koß, A. Patel, M. Helmberger, E. Szymańska, T. Gräber, and M. Fallon, "Hilti-Trimble-

Oxford Dataset: 360 Visual-Inertial Benchmark with Floor Plan Priors for SLAM and Localization," 2026. [Online]. Available: https://github.com/Hilti-Research/hilti-trimble-slam-challenge-2026

[2] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.

[3] A. Bikandi-Noya *et al.*, "BIM Informed Visual SLAM for Construction Monitoring," *arXiv:2509.13972*, 2025.