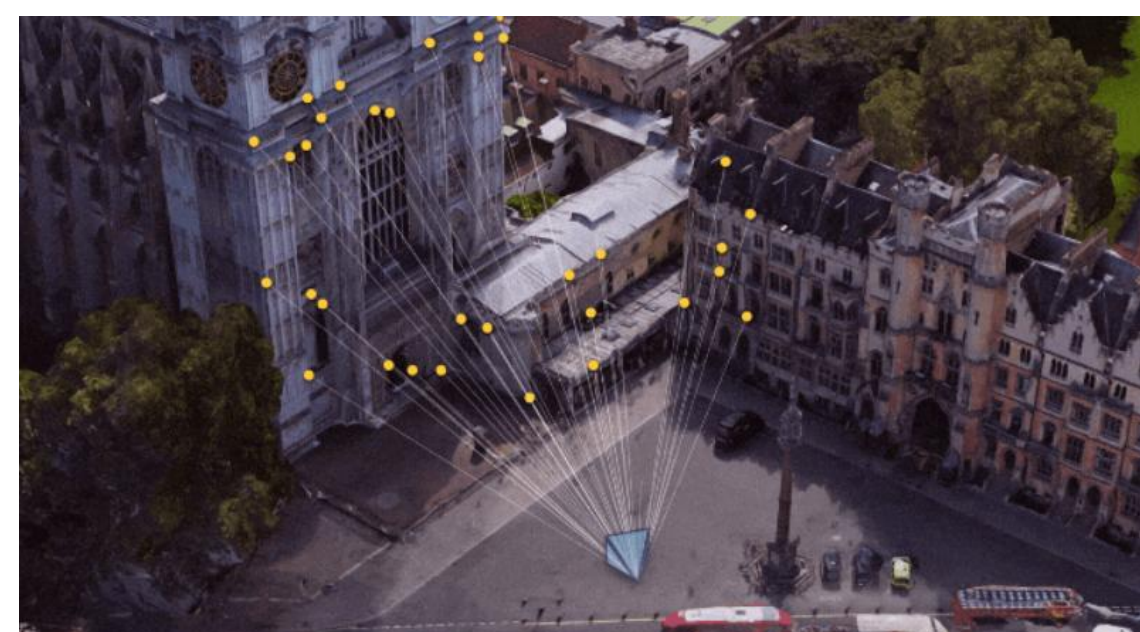
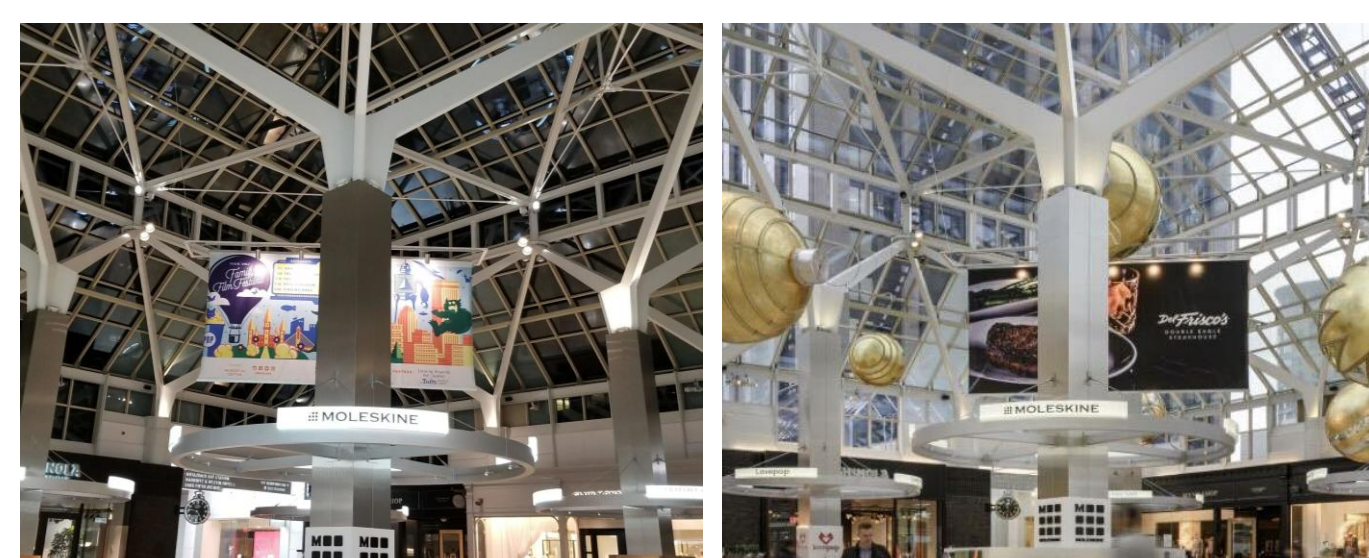


Background

Visual (re)localization is using camera to determine position and orientation, which requires construction of a map with visual features:



Cost of building 3D map



Outdated or transient map



Privacy & License
Wu et al., YOLO-SLAM

Overview

1. Capture image(s)



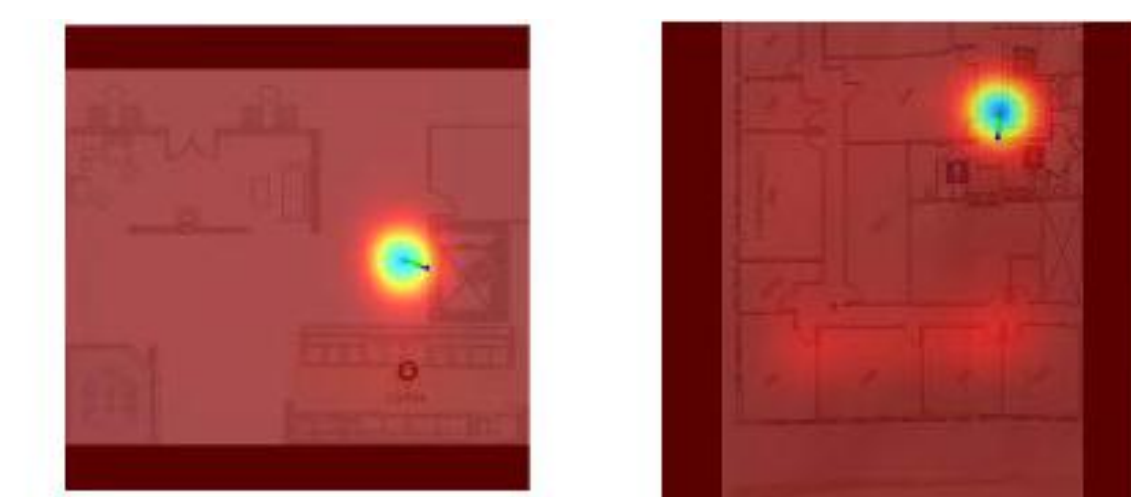
No calibration

2. Capture a floorplan



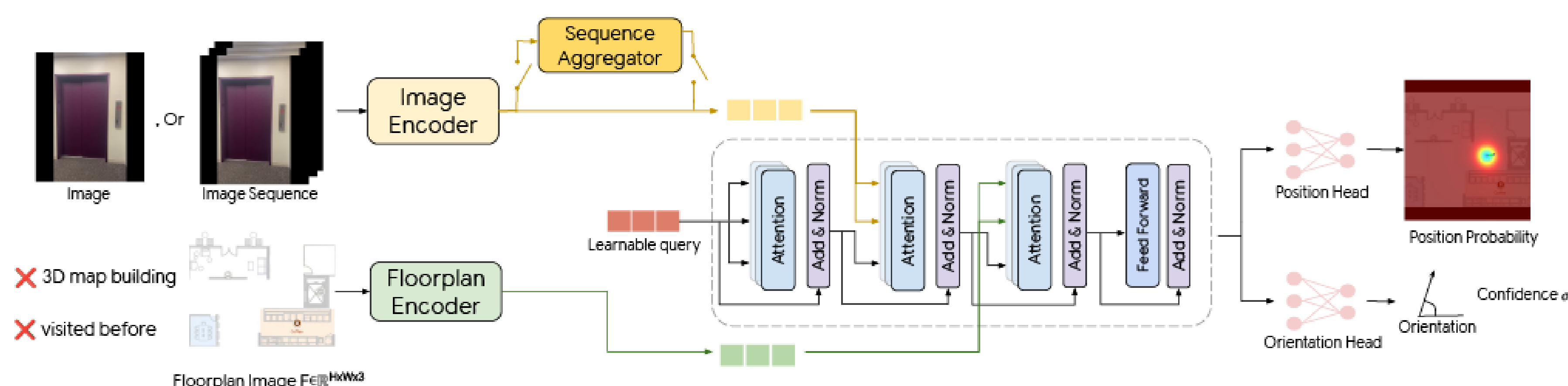
An image of floorplan

3. Inference localization



13.3 FPS / vram 2.04 GB

Method



We fuse image modality and floorplan modality to predict 3-DoF localization

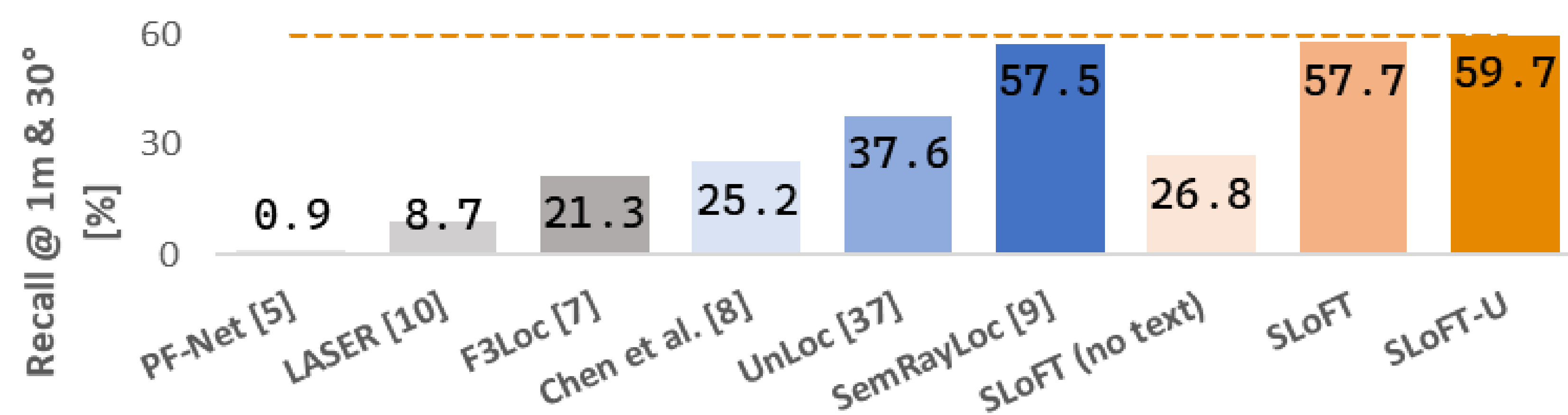
Training

Unified training with 1.2M+ images and 740K+ floorplans

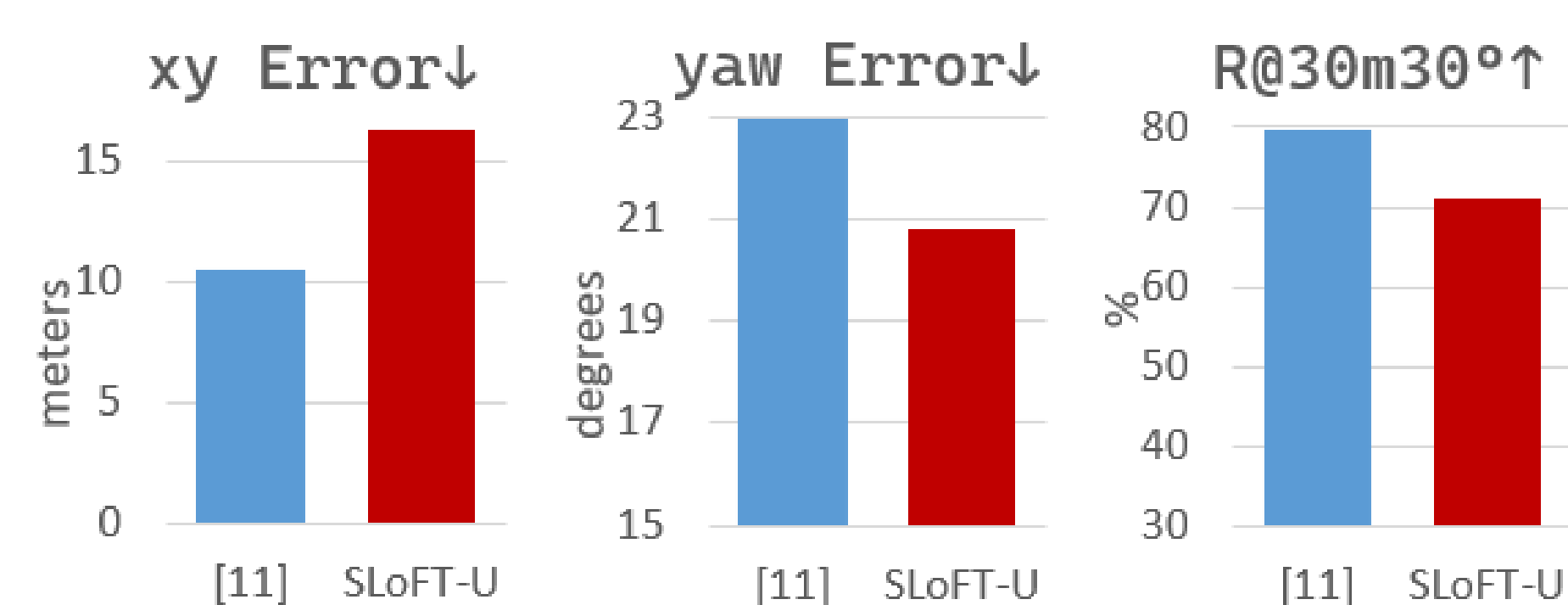
Dataset	Domain	Real?	Text?	Open vocab?	Symbols?
Structured3D[33]	House	✗	✓	✗	✗
MGL [11,32]	Outdoor	✓	✓	✗	✗
SLoFT-dataset	Office	✓	✓	✓	✓

Quantitative Comparisons

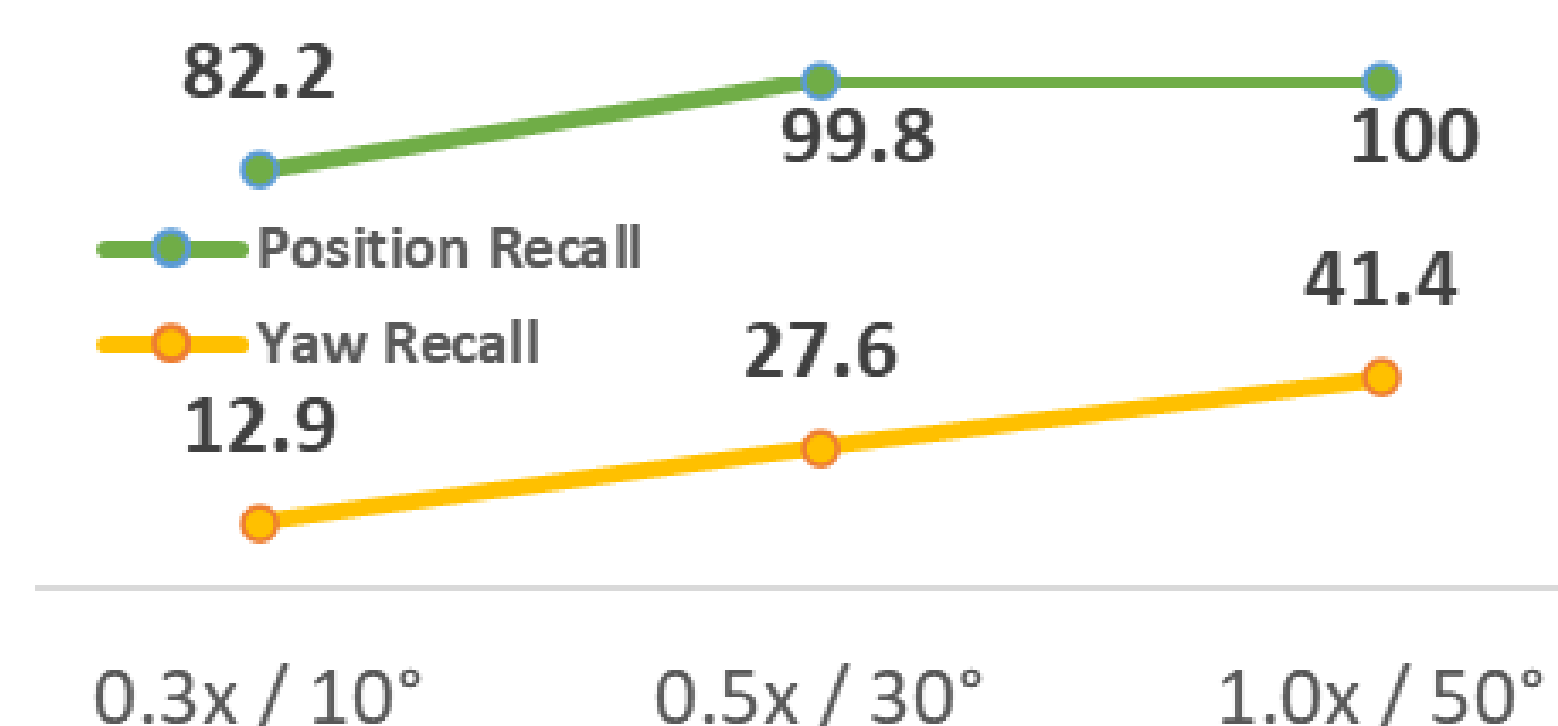
Structured3D dataset



(Outdoor) MGL dataset

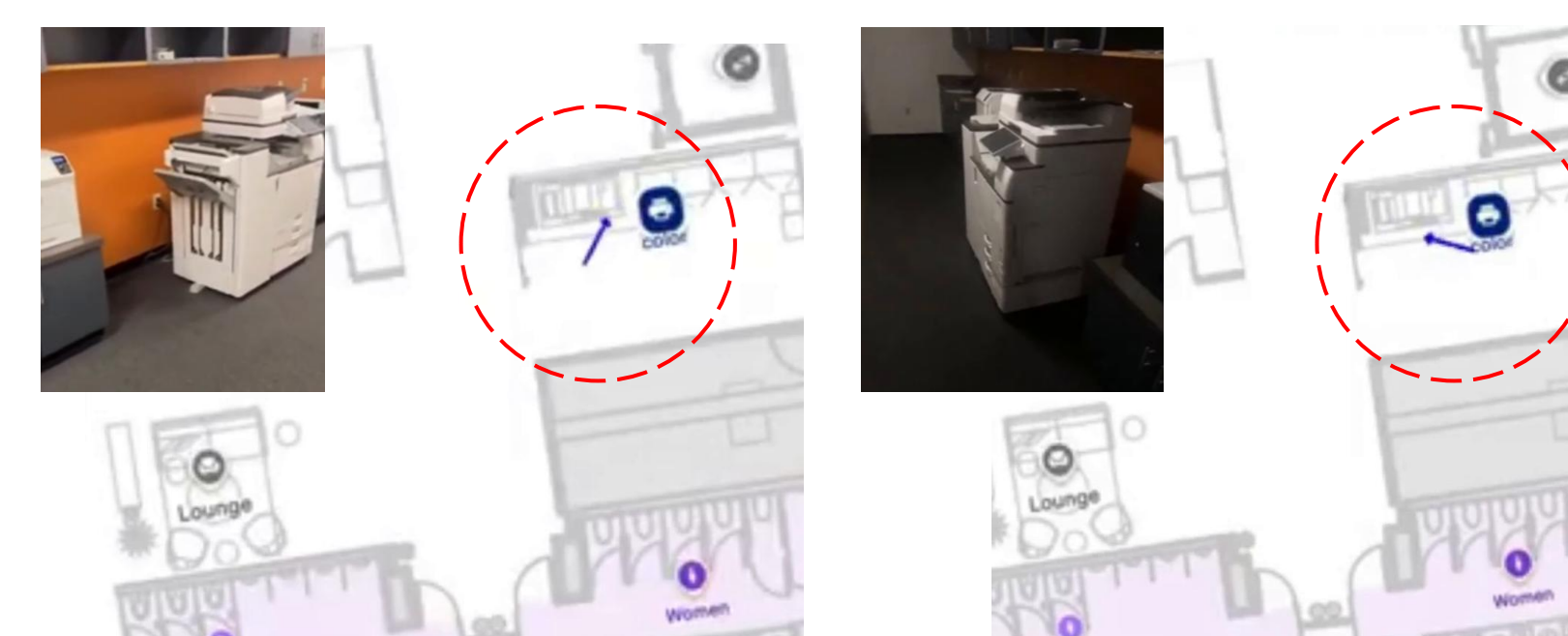


In-the-wild – Unseen buildings

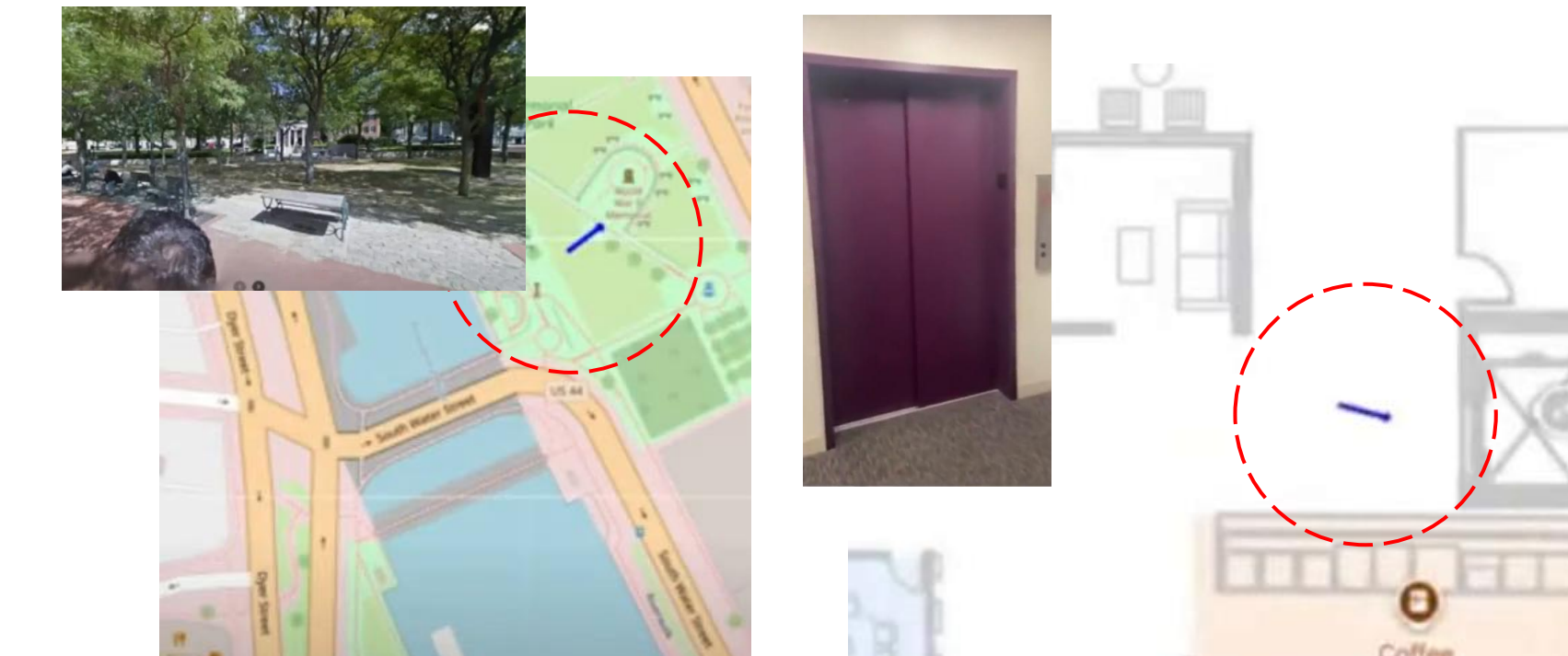


In-the-wild Results

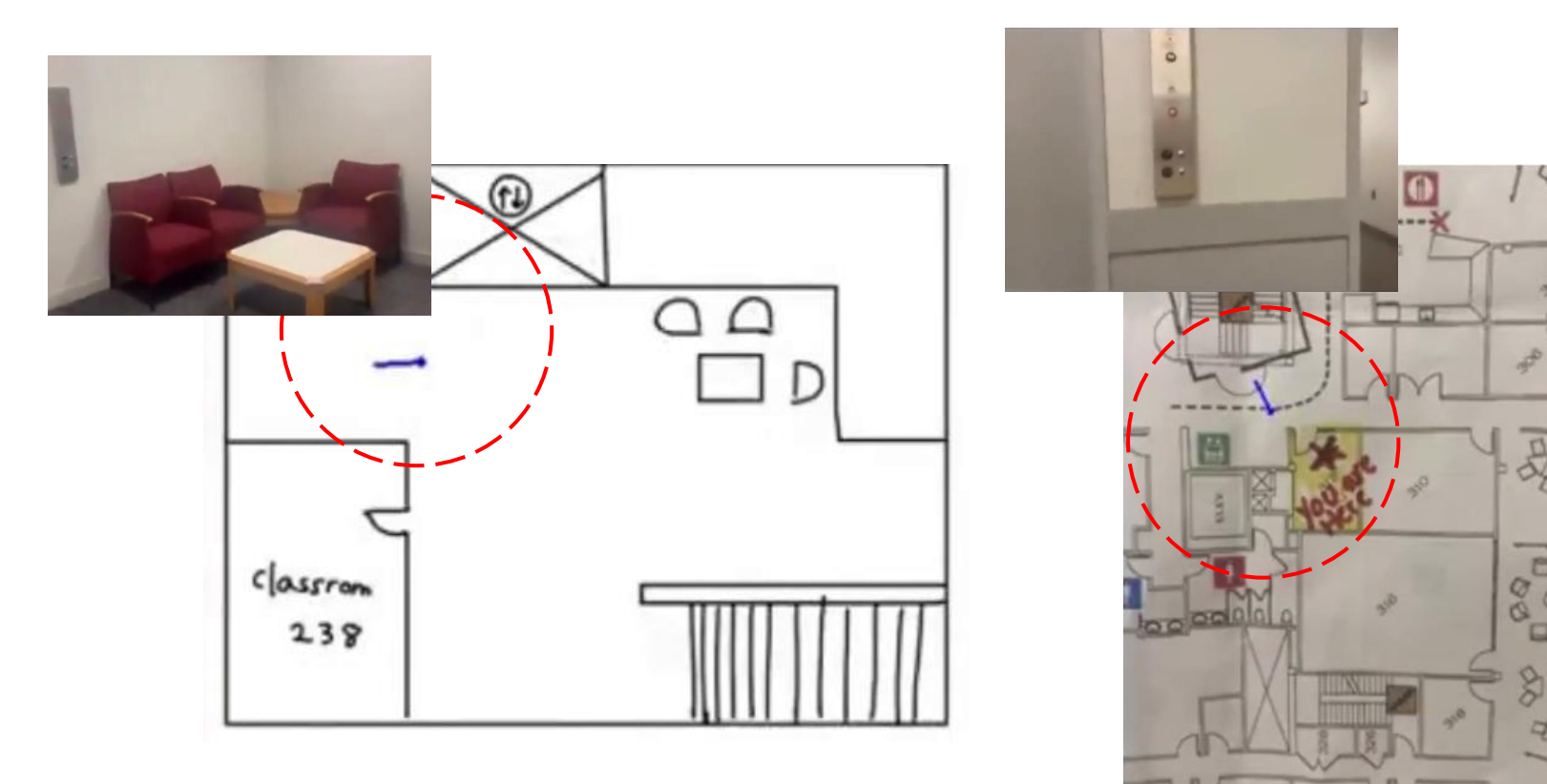
Varying lighting conditions



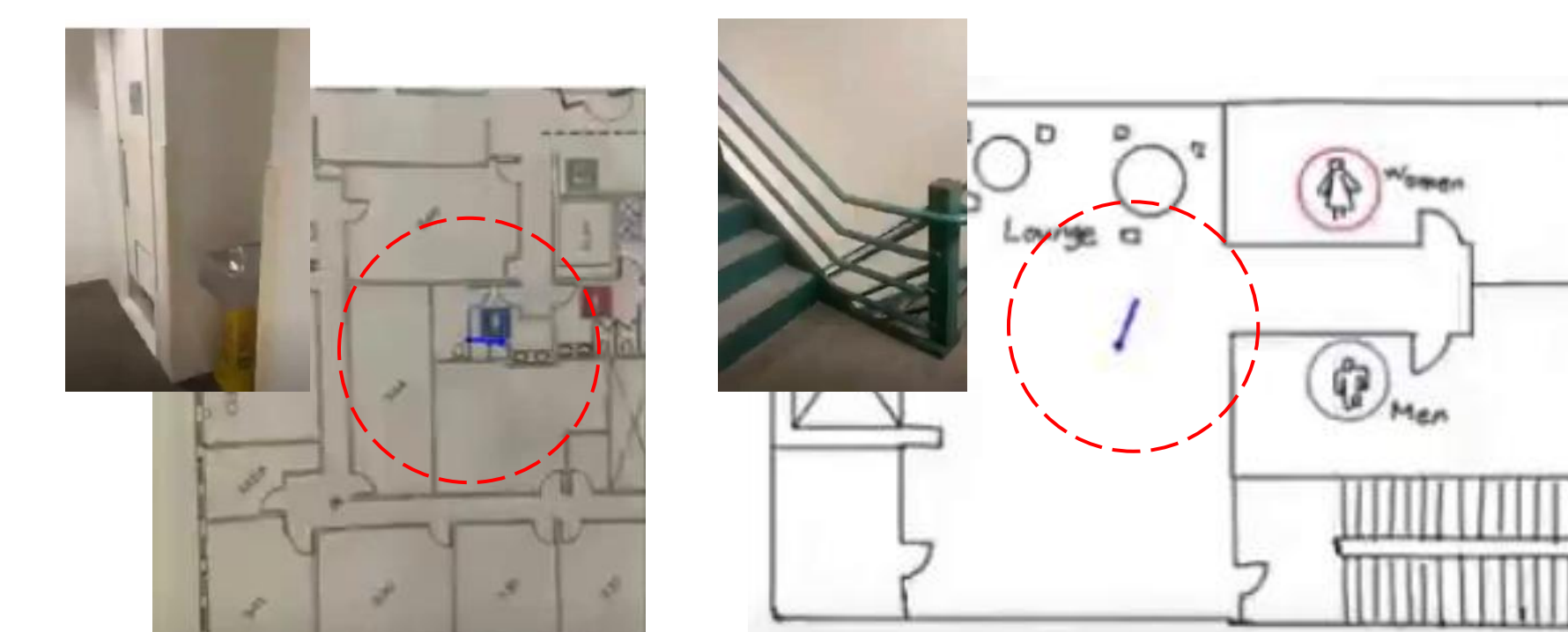
Indoor and outdoor



Different map styles



Failure cases



Doors cluttered in map

Rare case in dataset