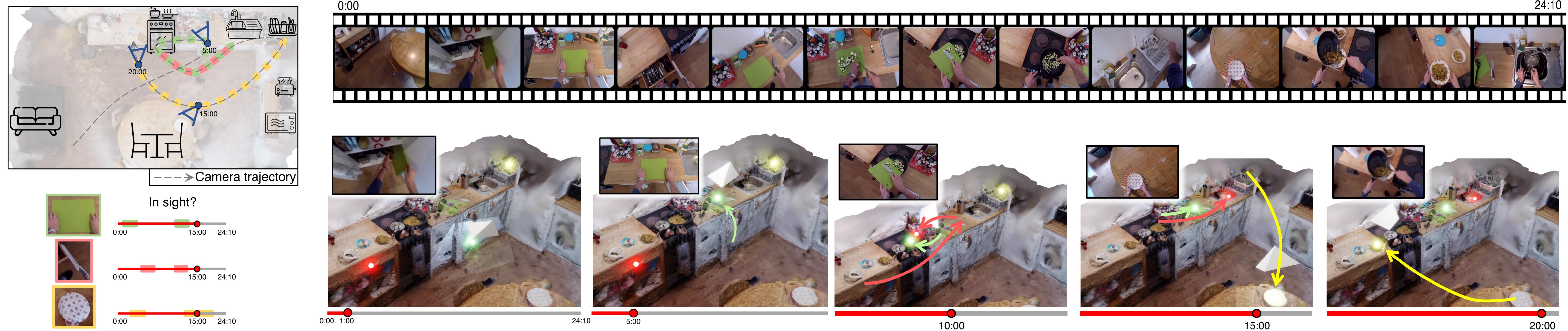


Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind

Chiara Plizzari Shubham Goel Toby Perrett Jacob Chalk Angjoo Kanazawa Dima Damen

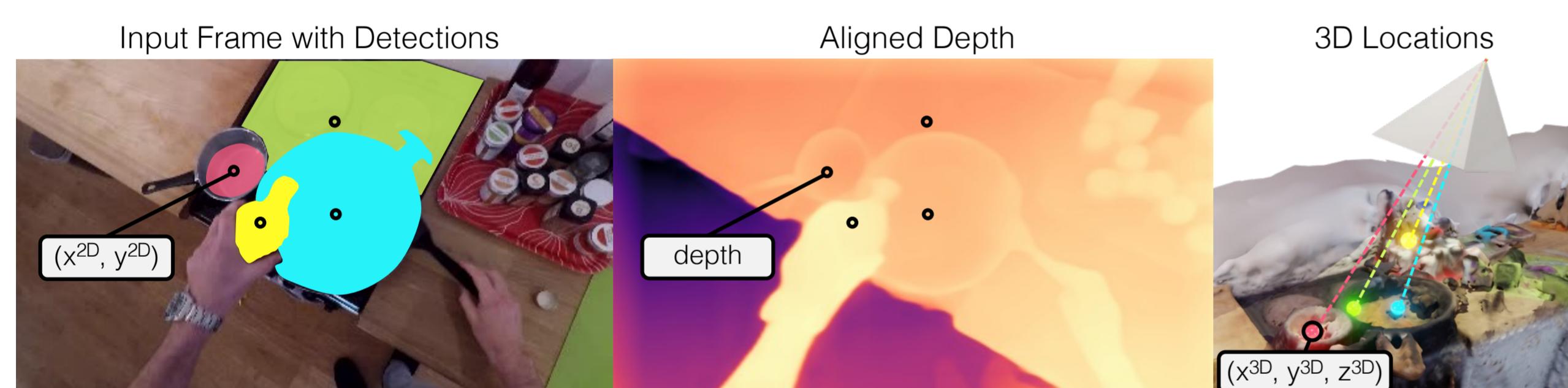
Motivation

- The ability to “know what is where” is an integral part of *spatial cognition*. It allows humans to build a mental map of the environment and dynamic objects.
- We introduce the task **Out of Sight, Not Out of Mind (OSNOM)** – maintaining the knowledge of where *all* objects are, even when absent from the egocentric video.
- We propose an effective approach that tracks objects in the world coordinate frame: **Lift, Match & Keep (LMK)**.



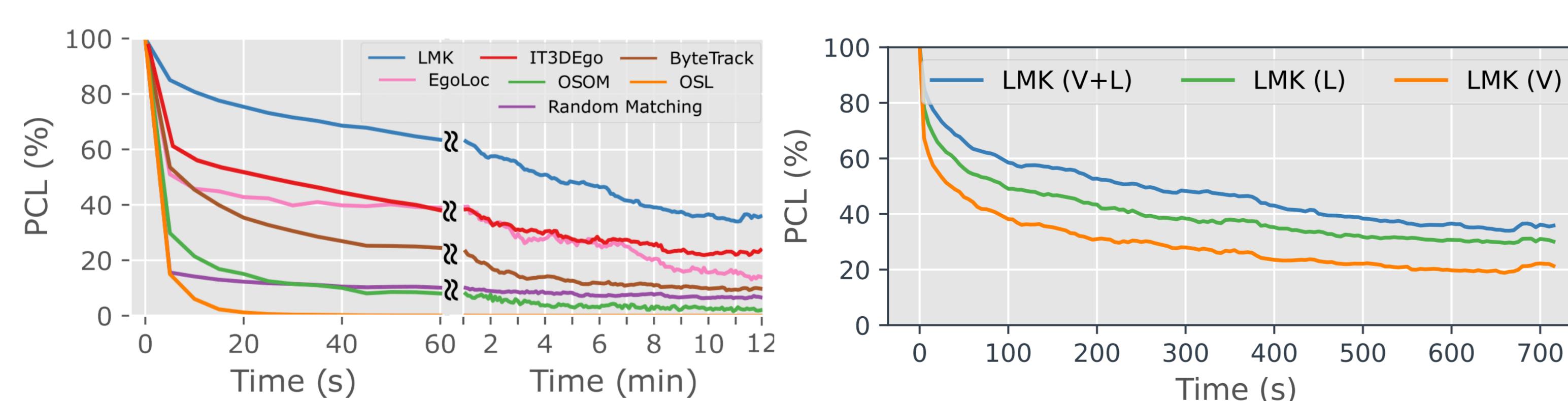
Method (LMK): Lift

- We lift 2D object detections (masks) to 3D using centroid of object masks and estimated mono depth in camera coordinate frame.
- We align the depth map to scene geometry so as to map these to world coordinate frames.

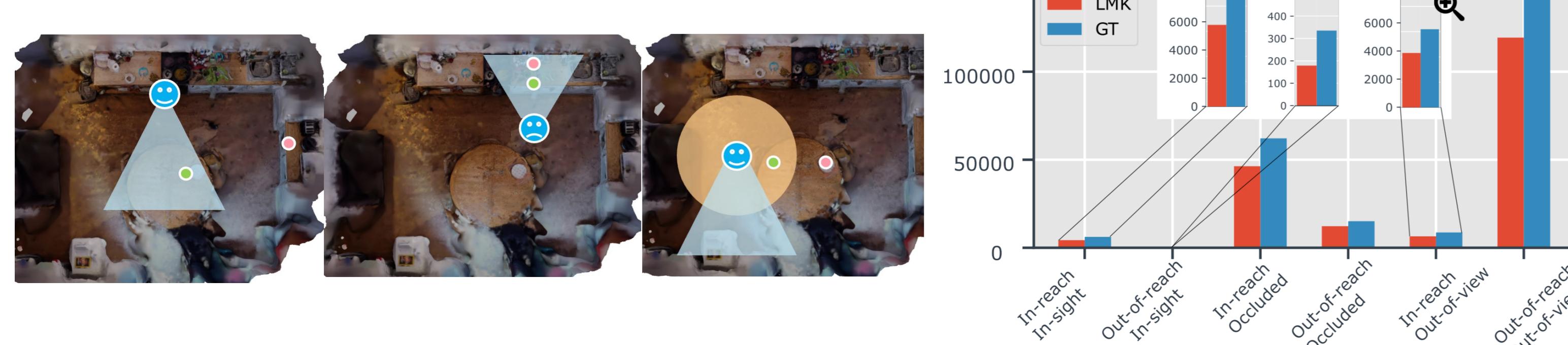


Lift, Match & Keep (LMK) Results

- We benchmark OSNOM on 100 videos from EPIC-KITCHENS, using the camera estimates from EPIC Fields [1].
- We introduce a new metric: *Percentage of Correct Locations (PCL)*.



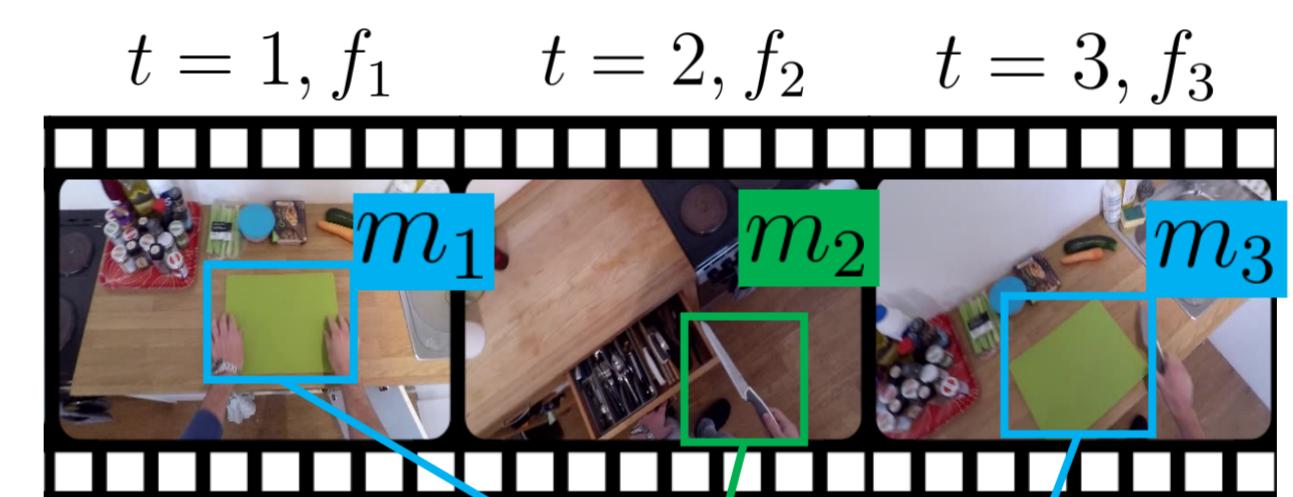
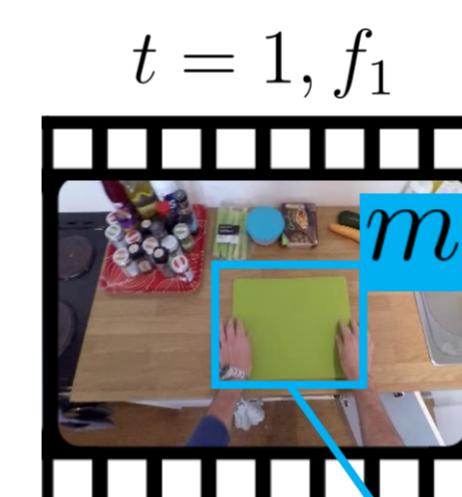
- After we Lift, Match and Keep (LMK) we can reason about object *visibility* and *positioning*: in-view vs out-of-view, in-sight vs out-of-sight (occluded) in-reach vs out-of-reach.



Method (LMK): Match & Keep

- Objects are tracked in 3D by *matching* visual and location features.

Track initialization



t : timestamp
 f_n : n-th frame
 m_n : n-th mask

n-th observation



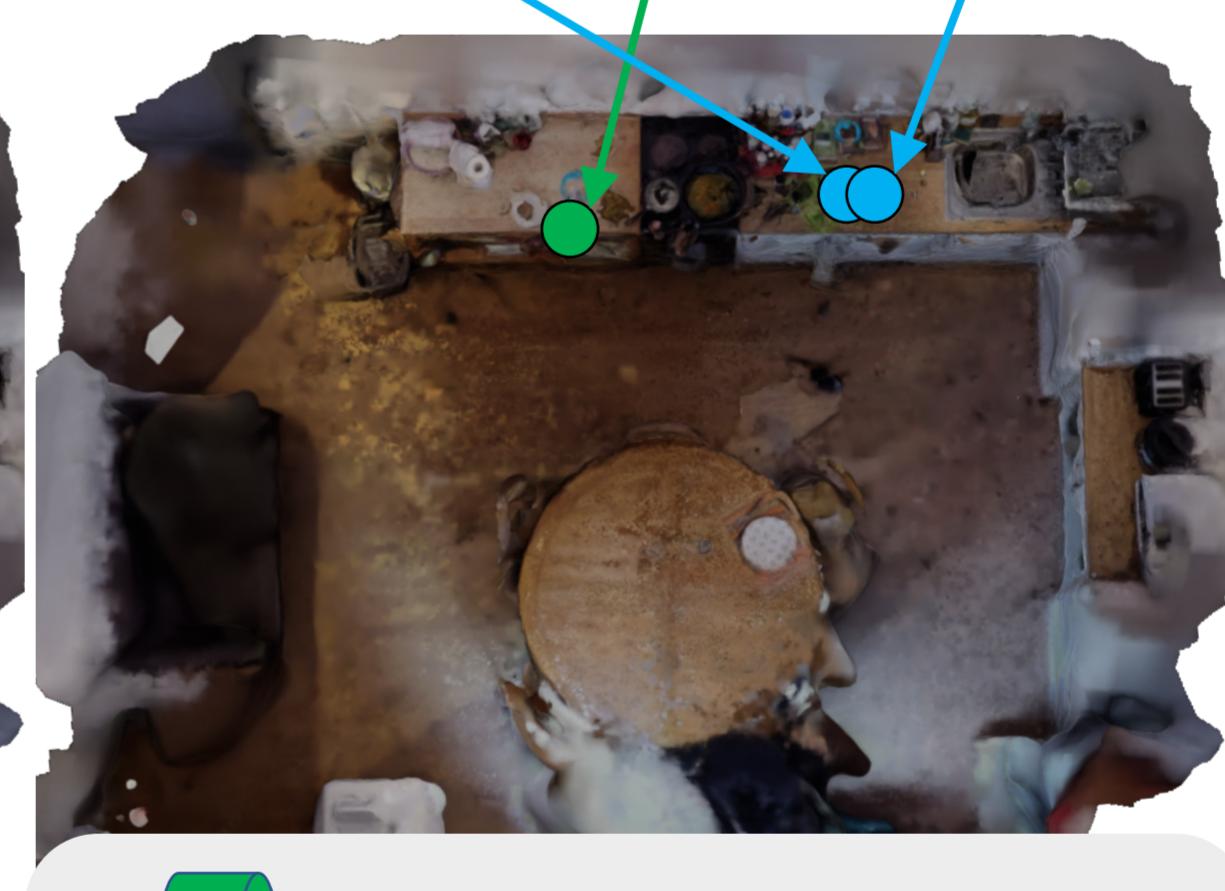
$o_n = (f_n, m_n)$

Visual Features

$$V_n = \Psi(f_n, m_n)$$

Location Features

$$\mathcal{L}_n = (x, y, z)$$



$$\mathcal{T}_{t=1}^i = (\mathcal{V}_1, \mathcal{L}_1)$$

$$\mathcal{T}_{t=2}^j = (\mathcal{V}_2, \mathcal{L}_2)$$

$$\mathcal{T}_{t=3}^i = (\mu(\mathcal{V}_3, \mathcal{V}_{t=1}^i), \mathcal{L}_3)$$

Qualitative Results

