A. Video examples

Sample results are present on the project's webpage: http://dimadamen.github.io/OSNOM. The video shows predicted object locations, over time, in 4 sampled clips from the evaluation EPIC-KITCHENS videos. We show the mesh of the environment, along with coloured neon dots representing the active objects that we lift and track in 3D. The videos also show the estimated camera position and direction throughout the video along with the corresponding egocentric footage.

In each case, the clip shows object locations predicted when they are in-sight, when they are out-of-view as well as when they are moving in-hand. Selected examples also show objects picked up / returned to fridge or cupboard highlighting the complexity of spatial cognition from egocentric videos.

B. Estimating error in the 3D projection

In Section 4.1, we estimate the error in 3D locations, through comparing projections of static objects from multiple viewpoint. Figure 3 in the paper presented the findings – showcasing that the mean error is 3.5cm with 96% of all errors within 10cm. We here describe the data used to report this figure.

We randomly selected 207,277 pairs of frames from our dataset, covering correspondences between 10 static objects across 5 different kitchens/environments. These were manually selected to find multiple frames with masks of the same object, at distinct times, and from different viewpoints. We avoid masks that are partially occluded by another object or by the camera's field-of-view (i.e. not fully in view) as these projections are likely to differ due to the occlusion of part of the mask. As the chosen pairs of masks showcase the same static object, their 3D locations should perfectly match. Any differences in their 3D location can be used to measure the error in the 3D projection, which we use as ground truth locations.

As the figure showcases, the error in our projections is within 10cm and well-within the threshold we use of 30cm. Recall that our threshold is chosen to reflect the cupboard width in standard kitchens. Estimating an object's location within 30cm implies we can position the object correctly within a cupboard.

C. Additional Ablations

Moved *vs*. Stationary objects. Section 3.3 also provides a definition of objects which have either moved significantly within the environment or remained relatively within a small section of the environment. We use a movement threshold of $\epsilon = 30\text{cm}$ to separate large from small motions. Figure 12 shows PCL results showing the objects that remain relatively stationary can be tracked on average 35% better than that of objects which have moved significantly within the space. Objects are more visually different after a move (*e.g.* different orientation or lighting).

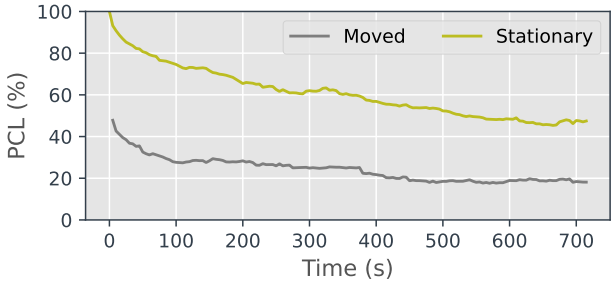


Figure 12. LMK Results for **Moved vs Stationary** objects with respect to the environment. We used a movement threshold of $\epsilon = 30\text{cm}$

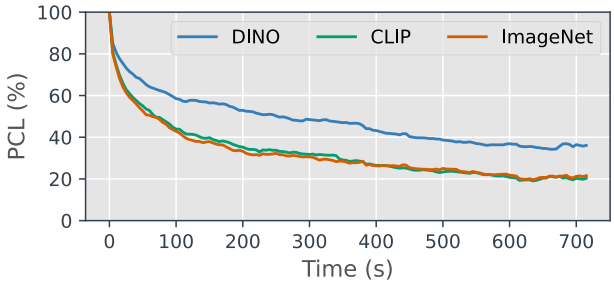


Figure 13. **Visual feature choice** of a DINO-v2, CLIP or ImageNet (ViT).

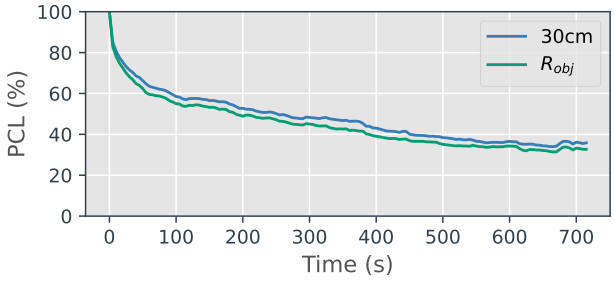
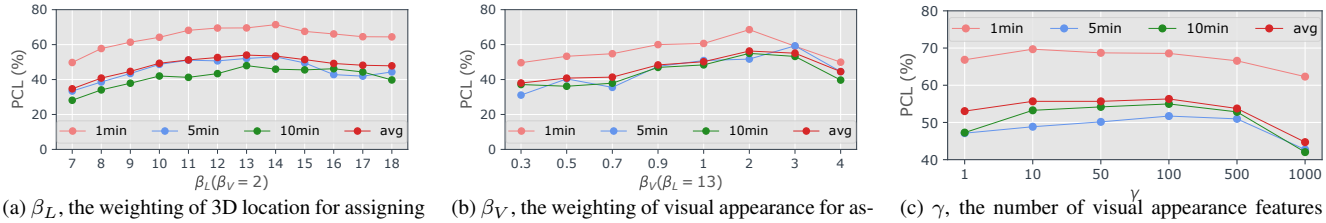


Figure 14. **Object radius.** LMK when approximating objects as spheres in 3D and using object radius for PCL threshold R.

Visual features. Our default feature extractor Φ is a ViT [10], pre-trained under the self-supervised DINO-v2 recipe [31]. We also compare to ViTs pre-trained on CLIP [34] and ImageNet [9] in Figure 13. DINO-v2 outperforms other approaches across all timescales, likely due to the pre-training tasks of CLIP (vision and language alignment) and ImageNet (image classification) being less suited to our requirement of reliable frame-to-frame visual similarity.

Object size. In our experiments, we use a fixed $R = 30cm$. As objects differ in size, one might argue that matching R to the object size is more reasonable. In Figure 14 we use an adapted R that matches the object dimension per example. Results are very similar to the default $R = 30cm$, showcas-



(a) β_L, the weighting of 3D location for assigning observations to tracks.

(b) β_V, the weighting of visual appearance for assigning observations to tracks.

(c) γ, the number of visual appearance features averaged to calculate track representation.

Figure 15. **Hyperparameter ablations** for LMK on the validation set. We choose the best average over 1, 5 and 10 minute sequence lengths.

ing that fixed versus dynamic R do not change the tracking capabilities.

Weighting visual appearance and location. LMK uses the hyperparameters β_L (Eq 4) and β_V (Eq 5) for relative weighting of visual and location similarities when assigning new observations to tracks. We select these based on best validation set performance averaged over timescales. Figure 15a shows validation set performance when fixing the chosen $\beta_V = 2$ and varying β_L. Figure 15b fixes $\beta_L = 13$ and varies β_V. Both hyperparameters are relatively stable, most likely due to the scaling by appropriate distributions (Cauchy and Exponential).

Track visual representation. Figure 15c ablates γ over the validation set – the number of recent features averaged for visual representation of a track. Best results are obtained with $\gamma = 100$, with worse results for small / large values of γ, with performance relatively stable even down to only one observation.

D. Failure cases

We identify two key reasons for failure cases for LMK. For clarity, we showcase each case separately – in Figure 16 and Figure 17. For each figure, we focus on a single object and show its predicted trajectory in green. Failure predictions are shown in red, where we plot the correct ground truth trajectory.

In Figure 16 we show cases where the track is lost for a limited time but is then correctly recovered. In the first row, the tin is correctly tracked for most of its trajectory, including when it is discarded in the bin. However, for a short duration, the predictions are incorrect (red dots). Similarly, in the second row, the knife is incorrectly predicted while occluded by the hand or occluded in hand. The last example shows failures in predicting the correct trajectory of the pot as it is filled with milk which changes its appearance. Coincidentally, it is moved out of the field of view. The matching then fails for both the appearance and the location. As the pot is emptied, its appearance matching is recovered towards the end of the track.

In Figure 17, we show failure cases of tracking that are not recovered. In the first example, the wooden spoon is assigned a new trajectory and the tracking continues using the new identity. This is similarly the case for the cutting board when it is moved to the cluttered sink.

Failures predominantly occur in cluttered scenarios, such as when slicing peppers with a knife in Figure 16, or mixing with a spoon in Figure 17. In these situations, the locations of multiple objects overlap, making the individual object's location less informative for matching.

E. Future Directions

We report the majority of our results using ground-truth masks out of the VISOR annotations. This allows us to evaluate the tracking from partial observations without accumulating detection errors. We find this decision to be reasonable as we focus on introducing and evaluating the task of Out of Sight, Not Out of Mind (OSNOM). In Fig 7, we ablate this by using an off-the-shelf semantic-free detector. The figure shows an expected drop in performance as noisy and incomplete detections are introduced. Improving performance using detection predictions is one of the future directions.

Another future direction is the expansion of OSNOM task to multiple videos, over time. In follow-up videos, the initial assumption of where objects are from previous sessions can be used as priors for OSNOM. Extending beyond a single video targets our ultimate goal of an assistive solution that is aware of where objects are, over hours and potentially days.

Figure 16. **Trajectory prediction - temporarily lost but recovered track.** Predicted trajectory of three objects in motion. Green neon dots show correctly predicted 3D positions over four frames with their corresponding camera views, and red neon dots show ground-truth trajectory where the prediction fails. The tracking momentarily fails, but subsequently, the object is accurately matched to a future observation.

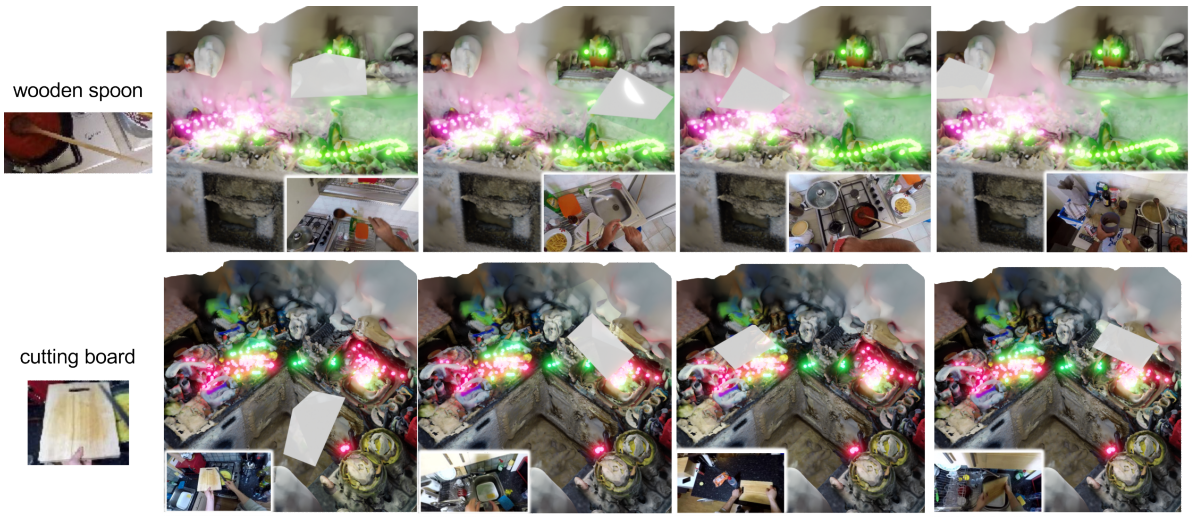


Figure 17. **Trajectory prediction - lost track.** Predicted trajectory of two objects in motion. Green neon dots show correctly predicted 3D positions over four frames with their corresponding camera views, and red neon dots show ground-truth trajectory where the prediction fails. The tracking fails and all subsequent predictions are assigned to a new track.