Appendix - AMEGO: Active Memory from long EGOcentric videos

Gabriele Goletto[1] Tushar Nagarajan[2] Giuseppe Averta[1]
Dima Damen[3]

[1] Politecnico di Torino, Italy [2] FAIR, Meta [3]University of Bristol, UK

https://gabrielegoletto.github.io/AMEGO/

In this appendix, we report additional details on AMEGO, the Active Memories Benchmark, additional results and visualisations. In Sec. 1, we include more visualisations of both queries and qualitative results across the complete range of questions in Active Memories Benchmark (AMB). We further detail AMB in Sec. 2. We then give more information on how we query AMEGO to obtain the answers required for the set of questions in AMB (Sec. 3). Next, in Sec. 4, we present additional ablations for AMEGO. Finally, in Sec. 5, we report the pseudocode version of the pipeline adopted in AMEGO.

1 Qualitative results

In Fig. 1 we show additional examples of sequencing questions, with all possible alternatives. AMEGO is able to correctly answer them, showcasing its sequencing capabilities.

On the project webpage we include videos depicting AMEGO representation on EPIC-KITCHENS videos. Similar to Fig. 1 in the main paper each row shows a different location spotted with our method. The white bar represents the temporal position of the frame depicted.

2 Active Memories Benchmark

2.1 Ground Truth

We combine annotations from EPIC-KITCHENS [1], VISOR [2], and EPIC Fields [3] to extract the ground truth used for creating our queries.

To obtain ground truths for locations, we filter out all frames with a high optical flow norm as these correspond to segments of video where the camera wearer is moving between locations. We then compute the intersection between the rays tracing from the camera through 5 pixels representing a crop of the image (four corners and centre pixel) and the mesh of the scene. For a frame size of 480×854, we selected the following pixels : (213, 240) as central-left, (640, 240) as central-right, (427, 120) as central-top, (427, 360) as central-bottom, and (427, 240) as the central frame. We then average the 3D points obtained, representing the locations where the subject focused for a period of time, indicating a potential interaction. The average is employed to reduce errors arising from noisy automatically extracted meshes.

Fig. 1: Examples of Q1-Q4. In green the right answer, correctly selected by AMEGO.

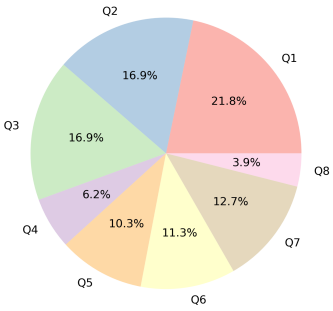
We then use hierarchical density-based clustering to obtain rough spatial clusters of the scene using the L2 distance among 3D locations as metric. Subsequently, we manually refined the clustering results to segment the videos into different functional activity-centric zones. For example, we differentiating a cooktop from the kitchen counter immediately adjacent to it, as they naturally afford different actions. We then find temporal segments corresponding to each location cluster. This approach enabled us to establish ground truths for temporal segments of locations.

To accurately capture object interactions' ground truths, we use action segments from EPIC-KITCHENS where brief interactions with the same object occurred. For example, we identify paired actions: 'open fridge' - 'close fridge' as well as 'pick plate' - 'put down plate'. By connecting these actions and finding the temporal extent between them, we can define the full interaction with objects. This approach allowed us to obtain the complete interaction interval, even when the camera wearer moved objects around the scene. We filtered out cases where different instances of the same objects appeared multiple times within a single video. Finally, we adopted VISOR masks to extract the visual queries for AMB.

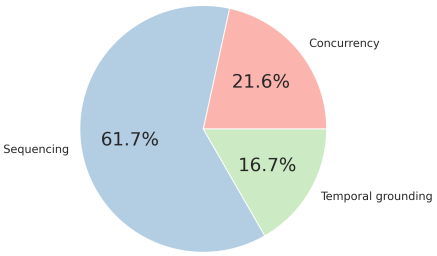
2.2 Query creation

The Active Memories Benchmark is a visual-only QA benchmark focused on the subject's interactions during long egocentric videos. One of the challenges is to select a visual representation for objects to form our visual query [VQ]. To address strong occlusion typical in egocentric vision, caused by the camera wearer's hands or other objects, we selected up to 3 different image patches for each object to form the query. These patches should be temporally distinct, to showcase different poses – we use a minimum of 0.5s between patches. Additionally, we select patches with minimal spatial overlap with bounding boxes of other active objects/hands in the same frame, to minimise occlusion. Similarly, for location images, we extracted frames with the lowest spatial overlap with active objects, so the location is present without many moving objects. For locations, we use location images with a minimum of 1s differences.

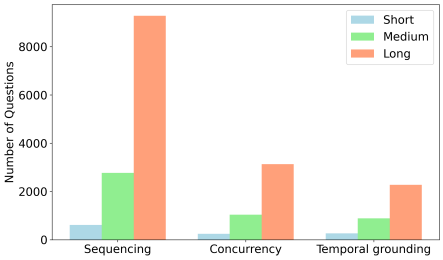
To create the Active Memories Benchmark we randomly sample 100 EPIC-KITCHENS [1] videos among the ones with both VISOR [2] masks and EPIC Fields [3] camera poses. We use the list of nouns from the narrations available in EPIC-KITCHENS, as an initial set of possible objects. We then filter out objects without corresponding VISOR masks, as we use these for spatial ground truth. We then generate the queries for all annotated objects/locations starting from the templates in Table 1 of the main paper. The alternative answers for each query have been generated using a rationale in a semi-automated process to increase the complexity of the benchmark.



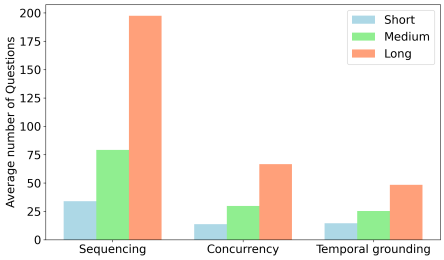
(a) Distribution of queries by type



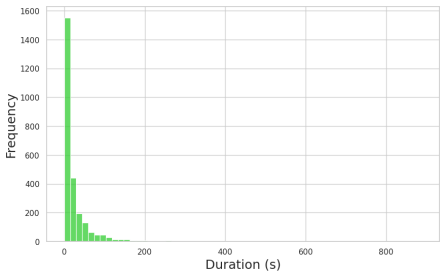
(b) Distribution of queries by reasoning level



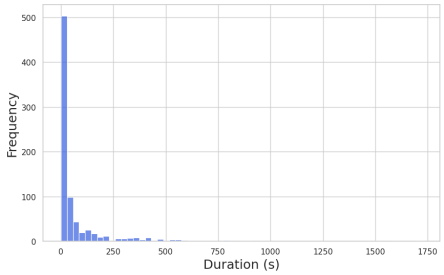
(c) Distribution of queries by video duration for each reasoning level



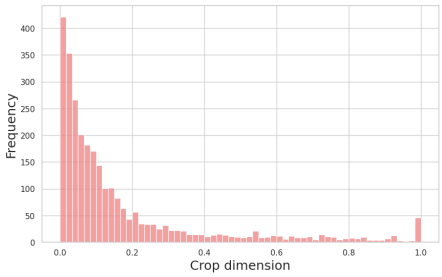
(d) Average number of queries per video by duration for each reasoning level



(e) Queries of type Q7 by duration



(f) Queries of type Q8 by duration



(g) Distribution of crop sizes of objects

(h) Wordcloud of most frequent queried objects

Fig. 2: Statistics of the Active Memories Benchmark questions

2.3 Statistics

In Figure 2, various statistics regarding the Active Memories Benchmark are presented. Specifically, the distribution per query type (Figure 2a), per reasoning level (Figure 2b), per video duration (Figures 2c and 2d), the interaction duration for queries of type Q7 (Figure 2e) and Q8 (Figure 2f), the distribution of crop sizes of the objects (relative to the frame dimension) in our benchmark (Figure 2g), and the frequency of queried nouns as a WordCloud (Figure 2h). Notably, sequencing queries constitute nearly two-thirds of the entire benchmark, reflecting their importance in understanding the temporal flow of object interactions in long videos, which is essential for higher-level understanding such as causal inference. Moreover, the number of questions increases with longer video durations (Fig. 2d), resulting in a benchmark tailored towards longer videos (the main focus of our work, see Fig. 2c). The duration of interactions exhibits a long-tailed distribution due to the fine-grained nature of queried objects and locations (Fig. 2e) and (Fig. 2f). Naturally, the most frequent objects in the dataset are also prominent in our questions (Fig. 2h). Smaller object crops are more frequently involved in our queries (Fig. 2g).

2.4 Benchmark comparison

Compared to other egocentric QA benchmarks, presented in Table 1, Active Memories Benchmark stands out due to its unique characteristics. It primarily emphasizes long egocentric videos, evident from the substantial average length of queried recordings. Similar to ReST [4], Active Memories Benchmark maintains a strong focus on vision, thereby mitigating potential language biases. However, unlike ReST, Active Memories Benchmark also incorporates the location dimension into its framework.

Table 1: Comparison of Active Memories Benchmark with other egocentric video QA datasets

Benchmarks	#Queries (K)	Avg. length (s)	Total hours	Vision focused
EgoVQA	0.5	2.2	0.3	
EgoSchema	5	180	253	
QAEgo4D	14.5	495	182	
ReST - ADL	185.7	1631	9	✓(Activity, Object, Time)
ReST - Ego4D	303.3	1104	92	✓(Activity, Object, Time)
AMB	20.5	1207	22.7	✓(Location, Object, Time)

3 Answering questions using AMEGO

Given the memory $\mathcal{E}$, our representation of the long egocentric video, querying it provides various ways to answer questions regarding interacting objects

and locations. Initially, we retrieve the closest representation of the query object/location, then apply the obvious logic to address various questions in the Active Memories Benchmark. Specifically:

- **Q1**: we match all objects in the sequences associated with each answer, assigning each image patch to an object ID q_{id}, among those in $\mathcal{O}$. Subsequently, we select the answer with the longest common subsequence calculated between the complete sequence of $\mathcal{O}$ and any of the answers;
- **Q2-3**: we match the query object with the track in $\mathcal{O}$ based on three criteria: (i) temporal proximity to t, (ii) containing the hand side specified in the question, and (iii) achieving a minimum similarity of 0.6. Once the matching track is identified, we simply extract the track after (Q2) or before (Q3) with the query hand side in it and search among the answers for the one with the highest similarity;
- **Q4**: we match the query object with the tracks in $\mathcal{O}$ and extract the corresponding object ID q_{id}. Using this ID, we identify the first/last location segment where it appeared in $\mathcal{E}$ and compare it with the answers. We select the answer with the highest similarity;
- **Q5**: we match the query object with the tracks in $\mathcal{O}$ and extract the corresponding object ID, q_{id}. Similarly, we match each image patch in the answers and extract the corresponding object IDs. Finally, we select the answer with the highest number of object IDs that are concurrent with q_{id} (i.e., overlapping temporal segments) in $\mathcal{E}$;
- **Q6**: similar to Q5, we match the query object with the tracks in $\mathcal{O}$ and extract the corresponding object ID, q_{id}. Then, we match each location patch in the answers and extract the corresponding location IDs. Finally, we select the answer with the highest number of location IDs that are concurrent with q_{id} (i.e., overlapping temporal segments) in $\mathcal{E}$;
- **Q7-8**: we match the query object or location to the tracks in $\mathcal{O}$ or $\mathcal{L}$ respectively. Then, we extract the corresponding instance and retrieve from $\mathcal{E}$ all temporal intervals where the instance was active. Finally, we select the answer with the highest average temporal Intersection over Union (IoU);

For each of the cases above, if two or more answers were found to be matching the representation, we select the answer randomly. Similarly, if no answer is found, a random answer is selected. We employ straightforward approaches to maintain focus on the strength of the representation rather than the querying method. The potential information extracted from $\mathcal{E}$ is solely constrained by the representation itself, and similar querying techniques can be readily implemented to address diverse, fine-grained questions about interactions in the long egocentric video $\mathcal{V}$.

4 Additional ablations

We present here additional ablation results on the clustering threshold to perform the assignment step for both objects (θ) and locations (τ) and for the IoU

Table 2: Ablation on IoU value

IoU	AIoU P ↑	AIoU GT ↑	ΔN → 0
0.3	0.19	0.37	226
0.5	0.20	0.41	210
0.7	0.18	0.32	168

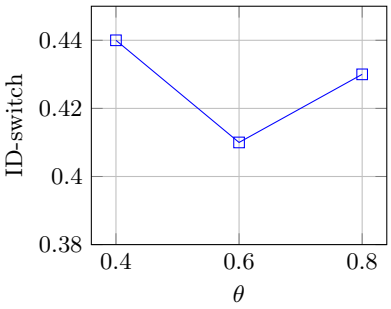


Fig. 3: Ablation on θ

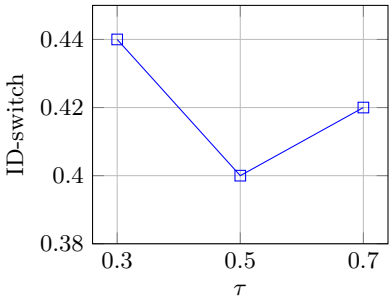


Fig. 4: Ablation on τ

threshold value adopted for spatial matching of HOI tracklets $\mathcal{O}$ with object detections $\mathcal{B}^o$. We performed the ablations on the two manually annotated videos described in Sec. 5.2 of the main paper. As it is important to evaluate clustering performance to ablate on θ and τ, we adopt another metric, ID-switch. It computes the average number of times a predicted segment changes its instance consecutively when it should not, i.e. when the ground truth object remains the same. Consequently, we want it to be as low as possible. Fig. 3 and Fig. 4 show the effect of the clustering threshold on ID-switch. Although clustering demonstrates stability across various thresholds, our selected threshold proves to be optimal for the two manually annotated videos. Similar deductions can be made from Tab. 2, where it is possible to notice that results do not change much depending on the IoU threshold chosen.

5 AMEGO pseudocode

For the sake of clarity we report here the pseudocode depicting the algorithms to build out interacting objects (Algorithm 1) and locations (Algorithm 2) representation.

Algorithm 1 Object interactions pipeline

Input:
  Frames $\{\mathcal{V}_t\}$
  HOI detector $\mathcal{D}$
  SOT tracker $\mathcal{T}$
  Similarity threshold θ

Output:
  Set of hand-object interaction tracklets $\mathcal{O}$

for each frame $\mathcal{V}_t$ **do**
  $\mathcal{B}_t^o, \mathcal{B}_t^h \leftarrow \mathcal{D}(\mathcal{V}_t)$ {Detect hands and objects}
  for each detection $(b^o, b^h) \in (\mathcal{B}_t^o, \mathcal{B}_t^h)$ **do**
    if new hand-object interaction (i.e. s_o detections in the last w_s frames) **then**
      Create new tracklet o_i
      Start SOT $\mathcal{T}_{o_i}$ for o_i
    end if
  end for
  for each tracklet o_i **do**
    Update the detections with $\mathcal{T}_{o_i}$
    if $\nexists b^o \in \mathcal{B}_t^o$ matching with o_i in the last e_o frames **and** $|\mathcal{B}_t^h| > 0$ **then**
      Mark o_i as complete
    end if
  end for
  for each completed tracklet o_i **do**
    Compute visual features $f(o_i)$ (Eqn. 1)
    Compute similarity $s(o_i, id_j)$ with existing instances in $\mathcal{O}$ (Eqn. 2)
    if maximum similarity $> \theta$ **then**
      Assign o_i to best matching instance id_j
    else
      Create new instance for o_i
    end if
    Store o_i in $\mathcal{O}$
  end for
end for

return $\mathcal{O}$

Algorithm 2 Location Segment pipeline

Input:
  Frames $\{\mathcal{V}_t\}$
  HOI detector $\mathcal{D}$
  Similarity threshold τ
Output: Set of location segments $\mathcal{L}$
for each frame $\mathcal{V}_t$ **do**
  $\mathcal{B}_t^o, \mathcal{B}_t^h \leftarrow \mathcal{D}(\mathcal{V}_t)$ {Detect hands and objects}
  Compute optical flow OpticalFlow$(\mathcal{V}_{t-1}, \mathcal{V}_t)$
  if location segment l_j is active **then**
    if high $|\text{OpticalFlow}(\mathcal{V}_{t-1}, \mathcal{V}_t)|$ **or** $|\mathcal{B}_t^h| = 0$ for e_l consecutive frames **then**
      Mark l_j as complete
    else
      Continue l_j
    end if
  else
    if low $|\text{OpticalFlow}(\mathcal{V}_{t-1}, \mathcal{V}_t)|$ **and** $|\mathcal{B}_t^h| > 0$ for s_l consecutive frames **then**
      Subject is interacting, start active location segment l_j
    end if
  end if
  for each completed segment l_j **do**
    Compute visual features $g(l_j)$
    Compute similarity $s(l_j, id_i)$ with existing instances in $\mathcal{L}$
    if maximum similarity $> \tau$ **then**
      Assign l_j to best matching instance id_i
    else
      Create new instance for l_j
    end if
    Store l_j in $\mathcal{L}$
  end for
end for
return $\mathcal{L}$

References

1. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV pp. 1–23 (2022)
2. Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations. In: NeurIPS (2022)
3. Tschernezki, V., Darkhalil, A., Zhu, Z., Fouhey, D., Laina, I., Larlus, D., Damen, D., Vedaldi, A.: Epic fields: Marrying 3d geometry and video understanding. In: NeurIPS (2024)
4. Yang, X., Chu, F.J., Feiszli, M., Goyal, R., Torresani, L., Tran, D.: Relational space-time query in long-form videos. In: CVPR (2023)