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/

Abstract. Egocentric videos provide a unique perspective into individuals' daily experiences, yet their unstructured nature presents challenges for perception. In this paper, we introduce AMEGO, a novel approach aimed at enhancing the comprehension of very-long egocentric videos. Inspired by the human's ability to maintain information from a single watching, AMEGO focuses on constructing a self-contained representations from one egocentric video, capturing key locations and object interactions. This representation is semantic-free and facilitates multiple queries without the need to reprocess the entire visual content. Additionally, to evaluate our understanding of very-long egocentric videos, we introduce the new Active Memories Benchmark (AMB), composed of more than 20K of highly challenging visual queries from EPIC-KITCHENS. These queries cover different levels of video reasoning (sequencing, concurrency and temporal grounding) to assess detailed video understanding capabilities. We showcase improved performance of AMEGO on AMB, surpassing other video QA baselines by a substantial margin.

Keywords: Long video understanding · Egocentric vision

1 Introduction

Episodic memory is a fundamental aspect of human cognition, which allows us to remember and recall our unique personal experiences [52]. Recently, there has been growing interest in leveraging first-person or *egocentric* videos to develop artificial episodic memory systems [11] that identify temporal segments from the video that contain answers to questions [38] or occurrences of objects [28,61] and activities [30, 69].

Critically, these approaches build representations of long videos from uniformly sampled frame or clip features, and then train a model to retrieve salient moments from the video using them. This has three drawbacks: (1) they are human activity agnostic — the simplistic uniform sampling of frames is done without an understanding of where the camera-wearer is, when object interactions occur, or what hand the camera-wearer uses, which are key parameters of human activity, (2) they rely on semantically labelled training data — explicitly training encoders to relate the query to the input video representations, and

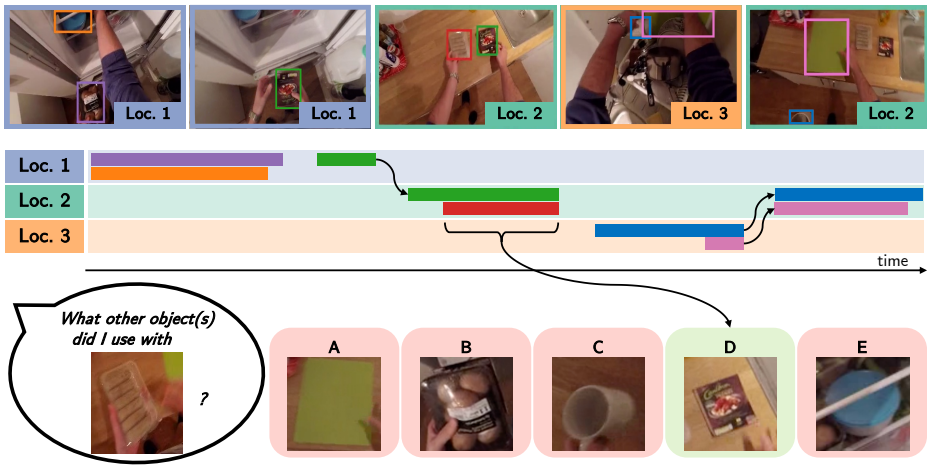

Fig. 1: AMEGO captures key locations and object interactions in a structured representation. In the each frame on top, the external border colour refers to a specific location in AMEGO while colours of objects define specific instances. AMEGO unlocks fine-grained long video understanding allowing multiple queries, such as the one depicted at the bottom of the figure, without reprocessing the long input video.

(3) they are difficult to interpret — the implicit representations do not directly reveal human activity leaving such approaches largely as *black boxes*.

To address these issues, we present AMEGO, an Active Memory of the EGOcentric video, which serves as an explicit, structured representation that captures both objects interacted with, locations visited, and the interplay between the two (see Fig. 1). Specifically, AMEGO is composed of (i) a collection of hand-object interaction (HOI) tracklets, which contain consistent interactions between the camera wearer and objects, and (ii) location segments, representing temporal intervals during which the camera wearer engages in activities within specific locations. Importantly, the tracklets and segments are built using visual perception models of human activity and motion rather than a naive sampling of frames. Such models capture information like "has an object-interaction begun or ended?", "is an interaction ongoing, despite hands not visible?", and so on, leading to representations that are directly tied to activities.

We populate our AMEGO memory following a three-step process: first, we identify the onset of object interactions; second, we detect the conclusion of ongoing interactions; and finally, we match concluded interactions to previously observed object or location instances. This online pipeline is preferred for efficiently storing and preserving only the relevant information, mirroring the way humans build episodic memories in everyday activities [51]. Each step uses off-the-shelf visual perception models, resulting in a training-free approach.

Noticeably, the resulting tracklets and segments are not associated with any fixed taxonomy of objects or locations, resulting in a semantic-free, queryable memory that can be used to answer questions about the video. By simply providing an image of an object or a location of interest, it is possible to access the

related information using feature matching. Thanks to its ability to use visual features in a semantic-free manner, AMEGO is more adept at distinguishing objects with subtle differences. This enhances its robustness and flexibility in capturing a wide range of interactions.

To evaluate AMEGO, we propose the Active Memories Benchmark (AMB) – composed of more than 20K visual question-answer pairs covering active objects, locations, and their interplay. We center the questions around 3 levels of reasoning, i.e. sequencing, concurrency, and temporal grounding. Notably, our benchmark is the first to tackle simultaneously interacting objects and locations. AMEGO achieves state of the art results on AMB, surpassing other common Video QA baselines by 12.7%. This performance underlines its capabilities across the three reasoning levels.

2 Related Works

Long video understanding. Long videos have gained significant interest recently, largely due to the emergence of large-scale egocentric datasets [5, 11, 12]. The task involves understanding videos lasting for minutes or even hours, leading to the creation of specialised models [14, 58]. Several question-answering benchmarks have been introduced to evaluate models' proficiency in understanding long videos [8,10,19,22,39,48,53,63,66]. Among these, the widely adopted benchmark EgoSchema [29] focuses on videos of up to 3 minutes in duration. Another benchmark, ReST [65], shares similarities with ours as it focuses on visual queries over long egocentric videos. However, it does not target locations and only emphasises object interactions. Different approaches have tackled long video understanding: some treat it as a natural language question answering task by first captioning the video and then using LLMs to answer queries [26,34,54,56,57,68]. Others integrate LLMs with a video encoder, leveraging the powerful comprehension and generation capabilities of LLMs [21, 36, 40, 45]. Our approach is similar to [9], proposing a structured representation of the video, but we specifically focus on interactions rather than indiscriminately memorising all the objects in the video.

Structured video representations. Various studies have explored methods for enhancing video representations by incorporating structured information. Contextual relationships have been a key focus, with numerous works investigating relationships between objects and actors [1,2,4,15,17,25,46,55], as well as among actions [3,14] using graph-based models. In the realm of egocentric vision, efforts have been made to construct structured representations of videos. For instance, [35] proposes grouping clips by activity threads, while [42] introduces egocentric scene graphs to capture interactions of the camera wearer. [32] focuses on constructing a human-centric representation of scenes by capturing the spatial locations of interactions, while [7] builds an allocentric top-down semantic scene representations, grounding the position of objects, from a video capturing a tour of the environment. Despite addressing various aspects of activities, these approaches do not capture the multiple dimensions inherent in egocentric videos — namely, object interactions, key locations, and their interplay.

Video summarisation. Another related task is video summarisation [27, 31, 41, 72] whose aim is to generate a shorter version of the video in the form of key frames or key shots. Egocentric summarisation approaches consider important people and objects [18], essential events [24] or aesthetic characteristics of key frames [59]. Some works have also proposed generating the summaries in an online fashion [23, 70] but do not target a structured representation of the video. [62] proposes a generic object finder, which automatically detects and clusters manipulated objects generating a timeline of the interactions. However, they do not exploit the temporal dimension proposing a system which is affected by noise coming from the detector. Another work which is related to ours is [60]. The method introduces a storyline representation for egocentric videos, summarising them based on actors, events, locations, and objects. It allows querying across dimensions using boolean operators. However, it mainly detects predefined attractions and supporting objects, which are visually distinct. Our work focuses on finer activities in cluttered scenes.

3 Method - AMEGO

Given a long and untrimmed egocentric video, we aim to capture the knowledge of active objects, key locations and their interplay using a unified structured representation. Such a representation must be *self-contained* — providing a full description of the camera-wearer's interactions with objects and locations — and *queryable* — as it should help retrieve temporal segments in the video indicating when an object was used, when a location was visited and their intersection (i.e. when an object was used in a specific location). In short, it is an Active Memory of the EGOcentric video, named hereinafter AMEGO.

We decompose the long egocentric video $\mathcal{V}$ into a set of hand-object interaction (HOI) tracklets ($\mathcal{O}$) and location segments ($\mathcal{L}$). Each **HOI tracklet** is a spatio-temporal representation of an object consistently interacting with at least one hand of the subject. It is characterised by spatio-temporal bounding boxes and their appearence features. Each **location** segment corresponds to the window of time where the camera-wearer *visits* a location to *perform* interactions, i.e we are interested in activity-centric zones or hot-spots for interactions.

Put together, the HOI tracklets and location segments form a memory of what objects the camera-wearer interacts with over time, in which locations, and how those objects are moved around the scene. This memory $\mathcal{E} = \{\mathcal{O}, \mathcal{L}\}$ is built online, eliminating the need for reprocessing past visual information, and then queried to answer a variety of questions about objects, locations and their interplay, as our experiments will show. Critically, our representations are *semantic-free* — they represent instances of objects and locations but are not tied to a fixed taxonomy of labels or a known vocabulary. They are tied to the visuals of objects and not to discrete categories, allowing a more fine-grained distinction.

In the following sections, we describe our pipeline to characterise and store object interactions (Sec. 3.1), to identify location segments (Sec. 3.2), and then to put them together to form our AMEGO representation. Finally, we describe how to query it to answer various questions (Sec. 3.3).

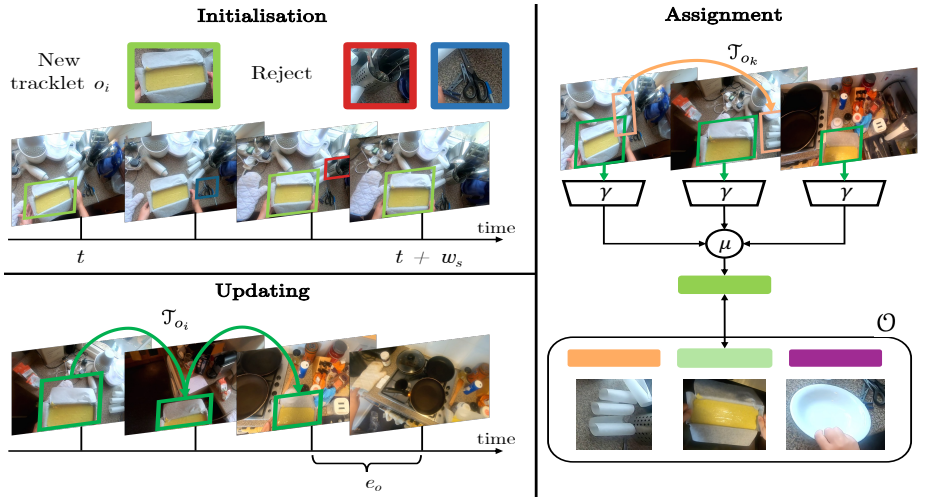

Fig. 2: We build $\mathcal{O}$ in an online manner, performing the 3 steps depicted at each frame of our video. (i) **Initialisation** We use consistent active object detections to generate new HOI tracklets. We thus discard noise resulting in sparse detections. (ii) **Updating** Once a new tracklet is initialised, we use a SOT tracker ($\mathcal{T}_{o_i}$) to update its detections even when hands go out of the field of view. We end the tracklet when there are e_o consecutive frames with a free hand or a distinctive new object interaction. (iii) **Assignment** Once a tracklet terminates, we assign it an object instance based on the similarity between its visual features wrt those in memory $\mathcal{O}$.

3.1 Object Interactions

We begin by characterising object interactions as HOI tracklets $\mathcal{O}$. Each tracklet $o_i \in \mathcal{O}$ is a tuple (t_s, t_e, b_t, h, id) where (t_s, t_e) are the start and end frames of the interaction, b_t is the sequence of bounding boxes representing the object in each frame, h is the hand side that performs the interaction (i.e., left or right), and id is the object instance associated to the tracklet.

We iteratively build $\mathcal{O}$. At each frame $\mathcal{V}_t$ we perform 3 main steps: (1) initialise possible new candidate HOI tracklets, (2) update the HOI tracklets that are active (i.e. corresponding to ongoing interactions), and (3) store the ones that terminate in the memory $\mathcal{E}$, and assign their corresponding object instance.

Initialisation We use a class-agnostic hand-object interaction detector [43], which provides a set of active object and hand bounding boxes denoted as $\mathcal{B}_t^o$ and $\mathcal{B}_t^h$ respectively. We initiate a new HOI tracklet o_i for each new hand-object interaction, defined as a tubelet comprising at least s_o bounding boxes exhibiting strong spatial overlap within a temporal window of w_s frames (Fig. 2, top-left). This spatio-temporal filtering allows us to account for noise as the result of hand-object detectors applied independently over frames. By leveraging the duration of natural hand-object interactions, we can reliably identify new active HOI tracklets, ensuring spatio-temporal consistency in the detections. The HOI tracklet o_i is now considered *active* and is added to $\mathcal{O}$.

For all subsequent frames, we calculate the intersection over union (IoU) for each object interacting with the same hand side. Matching bounding boxes over a threshold, θ, are assigned to the tracklet o_i. When bounding boxes cannot be assigned to the tracklet, it is considered complete.

Updating Next, we need to capture the entire duration of the interaction, and concurrently, capture all spatial occurrences of the object, performing interaction-aware tracking. While the frame-level HOI detectors are sufficient to identify new interactions, they are unable to reliably extend tracks over time where hands or objects exit the egocentric field of view. Instead, we use an off-the-shelf single-object tracker (SOT) [47] which can reliably track the object across the whole interaction. (Fig. 2, bottom-left).

Specifically, for each active object track o_i we initialise a SOT. We consider the track o_i completed if there are no associated detections $\mathcal{B}^o$ for e_o consecutive frames, while the hand h remains visible. This is because when the hand is out of view, it is likely to still be holding the object. This results in a spatio-temporal track $\mathcal{T}_{o_i}$ which tracks the object's position, but lacks information about the interaction itself.

At this point, o_i contains information about its temporal duration (start and end time) and spatial bounding boxes corresponding to the active object, by combining the strengths of frame-based HOI detection and the SOT.

Assignment and storing Finally, we match o_i to already seen object instances in our memory. Specifically, given the set of stored HOI tracklets $\mathcal{O}_t$ observed so far, and the set of running SOT tracks $\mathcal{T}_t$, we check whether o_i can be matched to an existing object instance or if we need to start a new one. To do this, we first compute the visual features of o_i:

$$f(o_i) = \frac{1}{|\mathcal{V}_{o_i}|} \sum_{k \in \mathcal{V}_{o_i}} \gamma(k, b_k^o) \tag{1}$$

where $\mathcal{V}_{o_i}$ is the set of frames associated with o_i, b_k^o is the detection for frame k and γ is a visual feature extractor (in our experiments, DINOv2 [33]). To match o_i with instances in $\mathcal{O}_t$, we use an online clustering approach based on $f(o_i)$. The similarity between o_i and a specific object instance id_j is computed as follows:

$$s(o_i, id_j) = \frac{1}{|\mathcal{O}_t \in id_j|} \sum_{\mathcal{O}_t \in id_j} < f_{\mathcal{O}_t}, f_{o_i} > \tag{2}$$

where $\mathcal{O}_t \in id_j$ are the HOI tracklets belonging to instance id_j and $< . >$ measures the cosine similarity. We assign o_i to the object instance id_j^* that maximizes Eqn. 2, and is above a specified threshold, θ. Note that if any tracker in $\mathcal{T}_t$ overlaps significantly with o_i, and the tracker confidence is higher than the maximum similarity above, then it is assigned to the tracker's instance. Otherwise, o_i is assigned to the corresponding instance id_j^*. If the maximum similarity is below the threshold, a new instance is created for o_i (Fig. 2, bottom).

At the end of this stage we associate $f(o_i)$ and the assigned instance to o_i, and store it into $\mathcal{E}$. We will refer to this *confirmed* tracklet as $\mathcal{O}_i$. It becomes part of AMEGO and can be consequently used in the querying process.

3.2 Location segments

We define the set of location segments $\mathcal{L}$ as the temporal segments when the subject is carrying out interactions at different activity-centric zones. As a subject may interact with multiple objects simultaneously but can only be present in one hot-spot at a time, each location segment $l_i \in \mathcal{L}$ is modelled as a temporal interval corresponding to the start and end of an interaction. Like object interactions, $\mathcal{L}$ is populated online, and in two steps as follows.

Temporal segmentation Given the egocentric frame $\mathcal{V}_t$ and the hand detections $\mathcal{B}_t^h$, to understand whether the hand is interacting with an object while being in a location, we compute the optical flow between $\mathcal{V}_{t-1}$ and $\mathcal{V}_t$ and check hand presence via $|\mathcal{B}_t^h| > 0$. We consider the subject carrying out a task if both optical flow has low norm and there is at least one detected hand. We used the criteria discussed above as proxies to determine whether the subject has paused (through low optical flow) and is actively interacting with the scene (through a detected hand). Similar to the process for determining HOI tracklets, we adopt temporal filtering and consider a location segment, l_j, to be active only if these two conditions are verified for a consecutive number of frames, s_l. Similarly, we terminate l_j when we observe a consecutive number of frames, e_l, with either optical flow norm above the threshold or absent of hand detections.

At the end of this stage, we have temporally defined l_j but we still need to match it to previous location segments at the same hot-spot.

Assignment and storing We utilise a visual feature extractor σ for locations, to compute average features for the location segment denoted as $g(l_j)$. Next, we calculate similarity scores between the stored location instances and l_j by computing the average cosine similarity. We assign l_j to the instance that maximizes the similarity beyond a specified threshold, τ. If the threshold is not met, a new instance is created. Finally, we pair $g(l_j)$ and the assigned instance to l_j, and store it into $\mathcal{E}$. We will refer to this *confirmed* location segment as $\mathcal{L}_j$.

3.3 Querying AMEGO representations

After processing the whole video we obtain our *AMEGO*: a complete set of HOI tracklets $\mathcal{O}$ and Location segments $\mathcal{L}$ (see Fig. 3). Utilising AMEGO, we can determine whether any object has been in use and if the person has interacted at any locations. We achieve this in a semantic-free retrieval manner. Given an object image, q_o, we first extract its visual features, $f(q_o)$, using γ. Subsequently, we assign it an object instance, q_{id}, based on the similarity between its visual features and those in $\mathcal{O}$. With the obtained q_{id}, we can query $\mathcal{E}$ to retrieve information about its interactions. For instance, by searching for all tracklets in $\mathcal{O} \in q_{id}$, and their interaction intervals (t_s, t_e), we can identify all the temporal segments when q_{id} has been used.

Similarly, using σ, we can match any input location image to $\mathcal{L}$ in an identical manner. Consequently, leveraging the common temporal dimension, we can understand where query objects have been used or what objects have been used

Fig. 3: An example of AMEGO on a long egocentric video depicting objects interacting with the left and right hand of the subject and the visited locations.

at the query location. This process allows us to answer any set of queries involving objects and locations without reprocessing the entire video. Inside $\mathcal{E}$, we encapsulate all the information about what occurred in the video. This transforms AMEGO into an active memory of the video that, regardless of queries, is aware of what interactions took place at any point in the video.

4 Active Memories Benchmark

We propose the Active Memories Benchmark (AMB) — a comprehensive framework to study the interaction between active objects, locations, and their interplay in long egocentric videos, which form key components of daily human activity. The benchmark consists of 20.5k queries covering various levels of reasoning. The queries take the form of multiple-choice questions ranging from simple questions about object use (e.g., What did I use with [VQ]? where [VQ] is a visual crop of an object). Given a set of [VA] visual answers, the task is to select the correct representation of an object that has been used at the same time as the object represented in [VQ]. Similarly, questions can be answered on the interplay between locations and objects, e.g. What locations did I use [VQ] in? The answer here would be a set of correct location representations [{LA, ...}]. Critically, each visual query of an object [VQ], visual object answer [VA], location query [LQ] or location answer [LA] are parameterised as *visual crops* [11, 65, 67] to mitigate the need for a fixed vocabulary or taxonomy resulting in biases associated with language. Forming a language-free benchmark avoids models that neglect visual data when answering the questions [16, 29, 64]. See Fig. 4 for a visual example.

4.1 Query Criteria

To construct our benchmark, we build a set of visual query templates that involve objects, locations, and their natural interplay (See Tab. 1). We structure our benchmark evaluation around three main reasoning levels, which serve as essential building blocks to enable higher-level activity understanding.

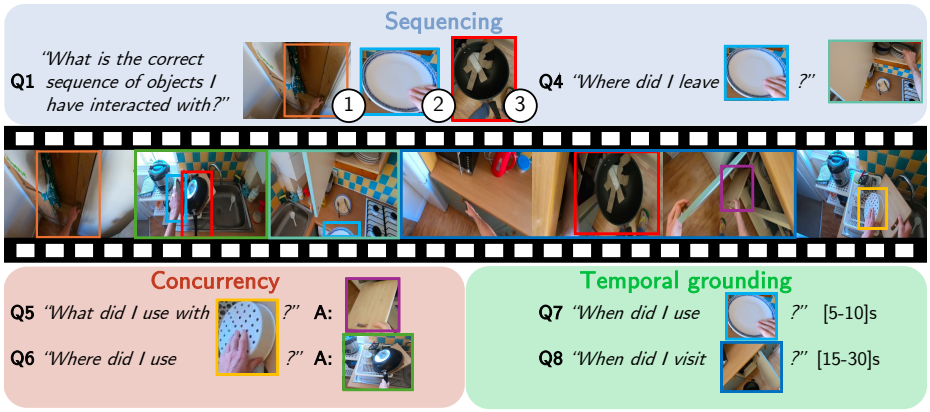

Fig. 4: Examples queries of Active Memories Benchmark on an egocentric video (in the middle). We build our benchmark around 3 different levels of reasoning, i.e. Sequencing, Concurrency and Temporal grounding.

Table 1: The question templates proposed in our benchmark, along with the corresponding required reasoning, dimensions, types of answers, and number of questions in Active Memories Benchmark. SQ, CO, and TG represent sequencing, concurrency, and temporal grounding respectively. [VQ] and [LQ] represent object and location crops, while O and L stand for object and location.

Reasoning	Query	Template	Dim.	Answer	Qs
SQ	Q1	What is the correct sequence of objects I have interacted with?	O	Obj. seqs	4464
	Q2	What did I use with the left/right hand *after* [VQ]?	O	Obj.	3466
	Q3	What did I use with the left/right hand *before* [VQ]?	O	Obj.	3466
	Q4	Where did I take/leave [VQ]?	O, L	Loc.	1266
CO	Q5	What did I use with [VQ]?	O	Obj. sets	2105
	Q6	Where did I use [VQ]?	O, L	Loc. sets	2320
TG	Q7	When did I use [VQ]?	O	Intervals	2614
	Q8	When did I visit [LQ]?	L	Intervals	809

Sequencing (SQ) questions assess the ability to discriminate the temporal order of events. For example, can the model order interactions in time and identify which object the subject used before or after using another object? These are captured by templates Q1-4.

Concurrency (CO) questions assess the ability to capture multiple interactions happening at the same time. For example, can the model reason about whether different objects have been used together (i.e. object-object concurrency), as well as whether an object interaction took place in a specific location (i.e. object-location concurrency)? These are captured by Q5-6.

Temporal grounding (TG) questions assess the model's ability to retrieve all intervals of interactions with an active object or a location within the long video. For example, can the model identify when a given object was used or a location was visited (i.e. the start and end time). These are captured by Q7-8.

These three aspects provide a holistic view of information stored in the active memory when observing the long video, serving as the foundational elements for task understanding and causal inference within procedural egocentric videos.

4.2 Benchmark Construction

We build our queries adopting the templates listed in Tab. 1 and express them as multiple-choice questions.

Egocentric Videos We construct our benchmark using 100 videos sourced from the EPIC-KITCHENS dataset [5]. This dataset is composed of long, unscripted egocentric recordings of participants performing daily living activities in a kitchen environment. On average, the selected videos are 14 minutes long. More specifically, 18 videos are shorter than 3 minutes, 35 videos are of medium length (between 3 and 10 minutes), 35 videos are long (10-30 minutes) and 12 are very long ($>$ 30 minutes). To define our ground truth, we leverage the publicly available dense camera poses from EPIC Fields [50] and active object masks from VISOR [6]. We segmented videos into activity-centric zones by leveraging camera positions from EPIC Fields to track the subject's attention on the scene. We merged EPIC-KITCHENS actions involving prolonged object usage and aligned VISOR masks with class semantics to obtain object bounding boxes. It is important to note that while we utilised these annotations for benchmark construction, our focus was solely on their application in evaluation. Additional details can be found in Supp.

Query generation We generate queries in a semi-automatic manner from our templates. AMB consists of 20.5K multiple choice question answer pairs. Each question consists of five possible options which are extracted semi-automatically from ground truths, to increase the challenge of these questions. In particular, we select candidate answers differently according to the type of question. For instance, for question Q6, options include locations visited immediately after the subject interacted with a specific object. Similarly, for questions Q3-4, options might include objects used with the opposite hand. This design makes AMB particularly challenging, demanding a detailed understanding of the events in the long video. Similarly for Q2-3, the query time t is set such that [VQ] is yet to be used – requiring the search for the interaction with [VQ] first before finding interactions before/after [VQ]. This systematic approach enabled us to create over 20,500 questions, of which 61.7% are sequencing questions, 21.6% concurrency questions, and 16.7% temporal grounding questions, see Tab. 1. Our dataset comprises 2614 object instances across videos and 809 activity-centric locations. On average, short videos ($<$ 3 mins) contain 62 questions, medium-length videos (3-10 mins) contain 134 questions, and long videos ($>$ 10 mins) have 313 questions.

5 Experiments

5.1 Experimental setup

Implementation details We use the hand-object interaction detector from [43] for identifying object-hand interactions at the frame level. Visual features of objects are extracted using the DINO-v2 pre-trained model [33] (γ), with resizing to 224×224 and evaluation on ViT-S and ViT-L versions. Object tracking during

interactions employ the EgoSTARK tracker [47] and we set $\theta = 0.6$, $w_s = 30$, $s_o = 20$, and $e_o = 20$.

For locations we use SWAG [44], σ, as the visual feature extractor, trained for image classification using weak supervision of hashtags. This model is currently state of the art in scene classification on Places-365 [71]. We evaluate on ViT-B and ViT-L versions with frames resized to 384×384 and 512×512. We estimate optical flow with the Flowformer model [13], and use a threshold of 2000 for the optical flow L2 norm. We set $s_l = e_l = 5$ and $\tau = 0.5$.

Baselines Our approach is the first able to create a complete representation of the long video which captures multifaced interacting elements. Prior works in this direction focused just on one specific dimension (e.g. locations [32] or activities [35]). SOT trackers would be able to track the object in the video but this would happen regardless of whether the object was interacted with. Consequently, we compare AMEGO against common baselines adopted for video QA on the proposed Active Memories Benchmark:

- **Semantic-free QA (SF-QA)** uses vision-language models, i.e. CLIP [37], to map the query, the video, and the answers into the same embedding space. This process involves extracting visual features from frames of the long video, query patches, and answers, while textual features are obtained from the question. The query embedding is generated by averaging the features from the video, patches, and question. Then, the similarity between this embedding and all answer embeddings is computed. The answer with the highest similarity score is selected.
- **SF-QA (obj)** is a variant of SF-QA, with visual features extracted also from active objects detected by [43].
- **Semantic QA (S-QA)** uses off-the-shelf captioners to generate a semantic summary of our video. We use the egocentric video captioner, i.e. LaViLa [68] at 1 fps, as in [68], and an image captioner, BLIP-2 [20], for generating captions of both the video and of the query patches. Increasing the captioning rate for video would only introduce redundancy in the resulting textual summary. Then, we input captions into an LLM and prompt it with the question. We use LLaMA-2-7B [49] because of its public availability, ensuring the reproducibility of our results. Due to the limited context window, we uniformly subsample textual summaries when they contain more than 4096 tokens.
- **Multi-round semantic QA (LLoVi)** [68] is similar to the previous one but it queries the LLM twice. First to summarise the video captions given the question, and then to answer the actual query based on the previously generated summary.

We evaluate AMEGO alongside the baselines in a zero-shot setting, measuring the accuracy over the queries provided in AMB.

5.2 Standalone performance

We first evaluate the effectiveness of the different AMEGO components against the ground truth. To do so, we manually annotate temporal interactions of objects and locations for two long videos: a 20-minute video from EPIC-KITCHENS

Table 2: Standalone evaluation for HOI tracklets (left) and location (right) segments

	AIoU P ↑	AIoU GT ↑	ΔN → 0
$s_o = 1$	0.08	0.49	3249
$e_o = 1$	0.13	0.34	537
Track w/o hand detections	0.19	0.38	222
No tracker	0.19	0.39	218
AMEGO	0.20	0.41	210

	AIoU P ↑	AIoU GT ↑	ΔN → 0
$s_l = e_l = 1$	0.14	0.48	163
No flow filter	0.35	0.35	-1
No hand filter	0.34	0.49	27
AMEGO	0.36	0.50	44

[5] (which is not included in AMB) and a 10-minute video from Ego4D [11]. Our annotations identify 22 distinct location instances and 67 different objects, allowing for a temporal comparison with the capabilities of AMEGO in defining interaction intervals.

We evaluate AMEGO using three metrics: (i) AIoU P, Average Intersection over Union between each predicted segment and its best-matching ground truth segment, indicating the precision of the predicted tracklets; (ii) AIoU GT, Average Intersection over Union between each temporal ground truth segment and its best-matching predicted segment, evaluating the recall of AMEGO; (iii) ΔN, the difference between the number of predicted and ground truth segments. Performance is improved when this number is closer to 0 ΔN → 0; Tab. 2 show the results for the both object and location interactions. In particular, it is possible to see the negative effect of noisy detections either at the beginning ($s_o = 1$) or at the end ($e_o = 1$) of the HOI tracklet. While AIoU GT is high for $s_o = 1$, this is due to the large number of segments predicted ($\Delta N > 3$K). Without using the hand detections, the tracking is stopped after e_o consecutive missing matches, regardless of hand presence. Accordingly, leveraging hand detections as a proxy to terminate the tracklet helps in detecting long interactions. Without using the tracker, the method performs worse as it is unable to track the object when the hands exits the field of view. Similar results for location segments show the importance of the various design decisions. It can be noticed how both flow and hand detection help to detect visited locations and, they are complementary.

5.3 Results on Active Memories Benchmark

To query AMEGO on AMB, we follow simple processes. As an example, to answer temporal grounding queries (Q7-8), we compare the query patch with instances in $\mathcal{E}$, as explained in Sec. 3.3, then extract the intervals in $\mathcal{E}$ corresponding to the matched instance. Additional details can be found in the Supplementary material. We report results on AMEGO- S, and AMEGO- L, depending on the size of the visual feature extractors adopted (ViT-S/B vs ViT-L).

Tab. 3 shows the main results on Active Memories Benchmark. All the baselines struggle to perform slightly better than random among the five answers. Particularly, it is noticeable that despite reaching high results on high-level understanding datasets [29], Semantic-QA approaches show limited understanding of fine-grained details on long videos. All the baselines show better results on concurrency-related questions (wrt the other reasoning proposed), which may hint at the fact that they might leverage training patterns, e.g. a pan often used

Table 3: Accuracy results (%) over the different queries of AMB. Best in **bold**.

Method	SQ				CO		TG		Total
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	
Random	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
SF-QA	13.7	21.6	22.5	26.8	22.1	31.9	23.7	26.2	22.0
SF-QA (obj)	13.1	23.4	22.6	23.2	21.7	26.1	23.8	25.2	21.2
S-QA (LaViLa)	20.9	20.6	21.2	24.6	24.9	27.1	21.4	22.6	22.4
S-QA (BLIP-2)	23.9	22.0	22.5	23.3	27.5	27.0	20.2	24.1	23.6
S-QA (LaViLa+BLIP-2)	22.8	22.2	21.4	22.6	25.1	26.1	21.4	24.5	22.9
LLoVi (LaViLa)	21.1	20.2	20.8	21.0	21.2	20.3	20.5	21.6	20.8
LLoVi (BLIP-2)	22.3	21.4	21.8	22.2	25.6	26.7	18.1	22.2	22.4
LLoVi (LaViLa+BLIP-2)	22.8	21.9	21.5	24.6	25.3	26.5	18.5	19.8	22.6
AMEGO - S	32.0	35.1	34.8	35.8	24.7	37.8	33.6	44.3	33.8
AMEGO - L	**33.7**	**36.3**	**37.2**	**38.3**	**27.6**	**44.3**	**34.7**	**48.9**	**36.3**

at the cooktop. The semantic-free QA baseline performs the worst, demonstrating that features by themselves, without a proper representation, are not enough. On average, BLIP-2 performs better on object-related queries. Indeed, differently from LaViLa, it has been trained on object-centric datasets and therefore shows superior capability to recognise them. Finally, we observe that multi-stage LLM pipelines, such as [68], perform worse than standard-QA. This likely depends on the fact that directly processing the textual summary reduces the amount of information at subsequent stages for correctly answering the query.

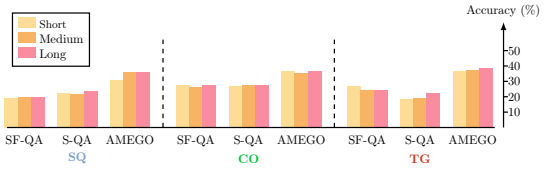

Fig. 5: Quantitative results depending on the temporal duration of the queried video.

AMEGO achieves good results on the whole set of queries, outperforming baselines by a large margin (+12.7%). It can be noticed that Q5 is the question where AMEGO struggles the most. This difficulty arises from current hand-object interaction detectors facing obstacles in predicting concurrent objects interacting with the same hand of the subject. Hence, despite our initialisation and tracking process allowing multiple objects to interact with the same hand, further improvements are needed in this regard.

Does video duration impact performance? We separately evaluate short (<3min), medium (3-10min) and long (>10min) videos. Fig. 5 compares the best performing semantic-free QA, semantic QA and AMEGO - L. In general, concurrency and temporal grounding questions are the ones that create more difficulty in long videos for SF-QA and AMEGO. This is reasonable as temporally locating objects and locations in longer videos is intuitively harder.

Qualitative results Fig. 6 shows two examples of concurrency and sequencing queries with the answers obtained querying AMEGO (in green) against the ones obtained via Semantic-QA (in red). AMEGO can understand the correct order

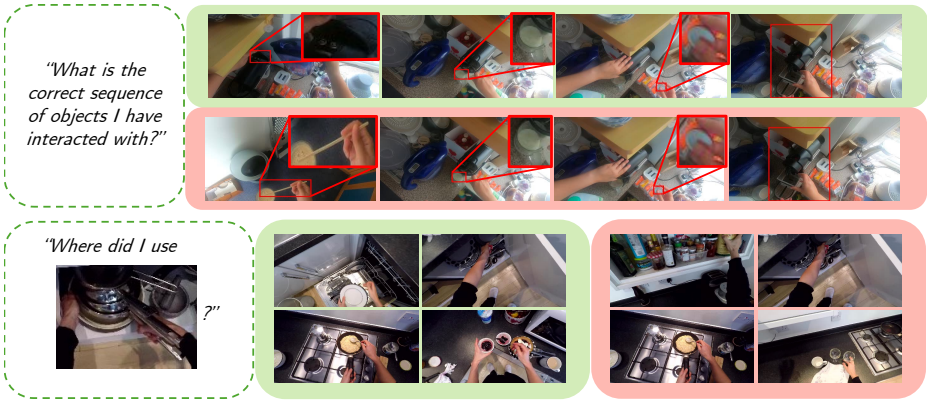

Fig. 6: Qualitative results are presented with sequencing and concurrency queries. Correct answers obtained from querying AMEGO have green background, while incorrect answers from Semantic-QA have red background.

of usage of items. Indeed it is possible to observe the steps performed by the camera wearer for preparing a coffee (upper part). The S-QA approach is not able to capture all fine-grained details in the video and is limited only to part of the sequence. In the bottom query, instead, it is possible to observe the training biases of LLMs preferring a cupboard (typically used to store a pan) rather than a washing machine.

6 Conclusion

In this work, we introduce AMEGO, an innovative Active Memory approach tailored for egocentric videos. By dynamically organising interactions and activities into a structured representation, in an online manner, AMEGO mimics the episodic memory cognition. Through semantic-free querying, AMEGO offers a powerful solution for efficient video comprehension without the need for exhaustive reprocessing.

We evaluate AMEGO on a newly proposed Active Memories Benchmark which underscores the effectiveness of AMEGO, showcasing its superior performance over common Video QA baselines. This highlights its ability to capture and represent intricate interactions within egocentric videos, paving the way for enhanced video understanding and analysis.

Acknowledgements G. Goletto is supported by PON "Ricerca e Innovazione" 2014-2020 – DM 1061/2021 funds and acknowledges travel support from ELISE (GA no 951847). G. Averta is supported by the project FAIR and Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. D. Damen is supported by EPSRC Visual AI EP/T028572/1 and UMPIRE EP/T004991/1.

We thank Chiara Plizzari for help in extracting the ground truths for AMB, Jian Ma for assistance in computing hand-object detections, and the members of the MaVi group for helpful discussions.

References

1. Arnab, A., Sun, C., Schmid, C.: Unified graph structured models for video understanding. In: ICCV (2021)
2. Baradel, F., Neverova, N., Wolf, C., Mille, J., Mori, G.: Object level visual reasoning in videos. In: ECCV (2018)
3. Brendel, W., Todorovic, S.: Learning spatiotemporal graphs of human activities. In: ICCV (2011)
4. Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., Yang, M.Y.: Spatial-temporal transformer for dynamic scene graph generation. In: ICCV (2021)
5. 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)
6. 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)
7. Datta, S., Dharur, S., Cartillier, V., Desai, R., Khanna, M., Batra, D., Parikh, D.: Episodic memory question answering. In: CVPR (2022)
8. Du, Y., Zhou, K., Huo, Y., Li, Y., Zhao, W.X., Lu, H., Zhao, Z., Wang, B., Chen, W., Wen, J.R.: Towards event-oriented long video understanding. arXiv preprint arXiv:2406.14129 (2024)
9. Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., Li, Q.: Videoagent: A memory-augmented multimodal agent for video understanding. In: ECCV (2024)
10. Fang, X., Mao, K., Duan, H., Zhao, X., Li, Y., Lin, D., Chen, K.: Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. arXiv preprint arXiv:2406.14515 (2024)
11. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)
12. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: CVPR (2024)
13. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: ECCV (2022)
14. Hussein, N., Gavves, E., Smeulders, A.W.: Videograph: Recognizing minutes-long human activities in videos. arXiv preprint arXiv:1905.05143 (2019)
15. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: CVPR (2016)
16. Jasani, B., Girdhar, R., Ramanan, D.: Are we asking the right questions in movieqa? In: ICCV Workshop (2019)
17. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: CVPR (2020)
18. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR (2012)
19. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: EMNLP (2018)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
21. Li, Y., Chen, X., Hu, B., Zhang, M.: Llms meet long video: Advancing long video comprehension with an interactive visual adapter in llms. arXiv preprint arXiv:2402.13546 (2024)

22. Li, Y., Chen, X., Hu, B., Wang, L., Shi, H., Zhang, M.: Videovista: A versatile benchmark for video understanding and reasoning. arXiv preprint arXiv:2406.11303 (2024)
23. Lin, Y.L., Morariu, V.I., Hsu, W.: Summarizing while recording: Context-based highlight detection for egocentric videos. In: ICCV Workshop (2015)
24. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: CVPR (2013)
25. Ma, C.Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., Graf, H.P.: Attend and interact: Higher-order object interactions for video understanding. In: CVPR (2018)
26. Ma, Z., Gou, C., Shi, H., Sun, B., Li, S., Rezatofighi, H., Cai, J.: Drvideo: Document retrieval based long video understanding. arXiv preprint arXiv:2406.12846 (2024)
27. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: CVPR (2017)
28. Mai, J., Hamdi, A., Giancola, S., Zhao, C., Ghanem, B.: Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In: ICCV (2023)
29. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: NeurIPS (2024)
30. Mavroudi, E., Afouras, T., Torresani, L.: Learning to ground instructional articles in videos through narrations. In: ICCV (2023)
31. Meena, P., Kumar, H., Yadav, S.K.: A review on video summarization techniques. Engineering Applications of Artificial Intelligence **118**, 105667 (2023)
32. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: CVPR (2020)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
34. Park, J., Ranasinghe, K., Kahatapitiya, K., Ryoo, W., Kim, D., Ryoo, M.S.: Too many frames, not all useful: Efficient strategies for long-form video qa. arXiv preprint arXiv:2406.09396 (2024)
35. Price, W., Vondrick, C., Damen, D.: Unweavenet: Unweaving activity stories. In: CVPR (2022)
36. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. arXiv preprint arXiv:2405.16009 (2024)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
38. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Naq: Leveraging narrations as queries to supervise episodic memory. In: CVPR (2023)
39. Rawal, R., Saifullah, K., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: Cinepile: A long video question answering dataset and benchmark. arXiv preprint arXiv:2405.08813 (2024)
40. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR (2024)
41. Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: ECCV (2018)
42. Rodin, I., Furnari, A., Min, K., Tripathi, S., Farinella, G.M.: Action scene graphs for long-form understanding of egocentric videos. In: CVPR (2024)
43. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: CVPR (2020)

44. Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R.P., Mahajan, D., Girshick, R., Dollár, P., Van Der Maaten, L.: Revisiting weakly supervised pre-training of visual perception models. In: CVPR (2022)
45. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR (2024)
46. Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., Schmid, C.: Actor-centric relation network. In: ECCV (2018)
47. Tang, H., Liang, K.J., Grauman, K., Feiszli, M., Wang, W.: Egotracks: A long-term egocentric visual object tracking dataset. In: NeurIPS (2024)
48. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: CVPR (2016)
49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
50. 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)
51. Tulving, E.: Episodic memory: From mind to brain. Annual review of psychology **53**(1), 1–25 (2002)
52. Tulving, E., et al.: Episodic and semantic memory. Organization of memory **1**(381-403), 1 (1972)
53. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Huang, S., Xu, B., Dong, Y., Ding, M., et al.: Lvbench: An extreme long video understanding benchmark. arXiv preprint arXiv:2406.08035 (2024)
54. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. arXiv preprint arXiv:2403.10517 (2024)
55. Wang, X., Gupta, A.: Videos as space-time region graphs. In: ECCV (2018)
56. Wang, Y., Yang, Y., Ren, M.: Lifelongmemory: Leveraging llms for answering queries in egocentric videos. arXiv preprint arXiv:2312.05269 (2023)
57. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. arXiv preprint arXiv:2405.19209 (2024)
58. Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C.: Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR (2022)
59. Xiong, B., Grauman, K.: Detecting snap points in egocentric video with a web photo prior. In: ECCV (2014)
60. Xiong, B., Kim, G., Sigal, L.: Storyline representation of egocentric videos with an applications to story-based search. In: ICCV (2015)
61. Xu, M., Li, Y., Fu, C.Y., Ghanem, B., Xiang, T., Pérez-Rúa, J.M.: Where is my wallet? modeling object proposal sets for egocentric visual query localization. In: CVPR (2023)
62. Yagi, T., Nishiyasu, T., Kawasaki, K., Matsuki, M., Sato, Y.: Go-finder: a registration-free wearable system for assisting users in finding lost objects via hand-held object discovery. In: International Conference on Intelligent User Interfaces (2021)
63. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: CVPR (2021)

64. Yang, J., Zhu, Y., Wang, Y., Yi, R., Zadeh, A., Morency, L.P.: What gives the answer away? question answering bias analysis on video qa datasets. arXiv preprint arXiv:2007.03626 (2020)
65. Yang, X., Chu, F.J., Feiszli, M., Goyal, R., Torresani, L., Tran, D.: Relational space-time query in long-form videos. In: CVPR (2023)
66. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Conference on Artificial Intelligence (2019)
67. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019)
68. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. arXiv preprint arXiv:2312.17235 (2023)
69. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: ECCV (2022)
70. Zhao, B., Xing, E.P.: Quasi real-time summarization for consumer videos. In: CVPR (2014)
71. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NeurIPS (2014)
72. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Conference on Artificial Intelligence (2018)