T-EVO: Tracking in Egovision for Online Visual Episodic Memory

Zaira Manigrasso[1], Antonio Finocchiaro[2], Davide Marana[1], Rosario Forte[2], Moritz Nottebaum[1], Matteo Dunnhofer[1,3], Giovanni Maria Farinella[2], Antonino Furnari[2], and Christian Micheloni[1]

[1] Machine Learning and Perception Lab, University of Udine, Italy
[2] Image Processing Laboratory, University of Catania, Italy
[3] Centre for Vision Research, York University, Canada

Abstract. Wearable assistants hold the promise of supporting humans in daily tasks, which requires a persistent awareness of the objects relevant to the user. However, existing methods typically operate on short video clips or rely on offline processing, limiting their capacity for long-term understanding. In contrast, humans are able to recognize specific object instances, recall previous interactions, and opportunistically retain useful spatial information. In this paper, we propose T-EVO (Tracking in Egovision for Online Visual episodic memory), a framework for online episodic memory that processes video streams online, storing compact, queryable object memories. T-EVO integrates an object discovery module, visual tracker, and a memory module to detect, track, and store spatio-temporal data of objects. Evaluated on Ego4D, T-EVO achieves an 81.9% success rate in the oracle configuration. However, its real-world performance drops sharply to 2.9%, highlighting significant limitations in detection and tracking capabilities. It enables fast, compact retrieval—cutting storage by 24× and retrieval time by 9×- demonstrating strong potential for real-world deployment in wearable devices.

Keywords: Egocentric tracking · Egocentric vision · Episodic memory

1 Introduction

Wearable devices equipped with cameras, such as smart glasses and augmented reality headsets, have the ability to perceive the world from the user's point of view. This egocentric perspective makes them ideal tools for personalized applications, capable of continuously assist the user. By keeping track of observed objects, these devices can alleviate cognitive overload and support the execution of daily tasks [21], ultimately assisting in the recall of episodic memories [26] across a wide range of scenarios. So far, models designed to analyze user-object interactions have largely relied on class-agnostic approaches [24]. Additionally, they have often been limited to short video clips or static frames, offering only a brief and fragmented understanding of interactions. Another common limitation is the offline processing paradigm, where algorithms retain a vast video archive

and can reprocess it at any time to search for specific objects [12, 13, 27, 32, 33]. However, humans perceive and remember object interactions in a fundamentally different way. Rather than thinking in terms of broad categories, we tend to focus on specific instances—recognizing a specific knife rather than just any knife. Our understanding is also long-term, allowing us to recall where and when we previously used an object. Moreover, the natural behavior is often opportunistic: we might unconsciously remember the location of an object that could be useful for a future task, even if it wasn't relevant at the time. A step toward enabling such cognitive abilities is the Visual Queries 2D (VQ2D) task, introduced as part of the Episodic Memory (EM) benchmark within the Ego4D dataset. The VQ2D task tackles a key challenge in egocentric perception: given an image crop of a queried object and a video clip, the objective is to determine the object's most recent occurrence by identifying its precise spatio-temporal position within the video. This approach provides a structured way to model how objects reappear over time from a first-person perspective, making it a foundational step toward systems that can track interactions in a long-term, memory-driven fashion. Inspired by this natural human ability to process interactions fluidly and contextually, our goal is to develop a system capable of real-time video processing without the need to store vast amounts of data for later re-analysis. This system would identify relevant objects, track their movement over time using an instance-based approach, and create a compact memory that summarizes the spatio-temporal evolution of important objects. These stored insights could then be leveraged to assist the user in future tasks, enhancing the effectiveness and usability of wearable devices. For instance, a system capable of detecting and tracking all relevant objects from an egocentric perspective could serve as a powerful memory assistant. Such a system could remind users where they left specific objects (e.g., "Where did I leave my keys?") and unlock new possibilities, particularly in the healthcare sector, where wearable systems could be used for cognitive monitoring and personalized training programs. The ability to comprehend user interactions from a long-term perspective represents a crucial step toward the development of intelligent systems that can adapt to individual needs, ultimately making technology more intuitive and responsive.

In sum, this paper presents a comprehensive online framework for the Tracking in Egovision for Online Visual episodic memory task (T-EVO). Our key contributions include: (1) the introduction of online VQ2D, an online approach for episodic memory retrieval. Figure 1 provides a schematic comparison of online and offline VQ2D; (2) the proposal of a novel framework that combines continuous object discovery, tracking, memory updating, and querying.

2 Related Work

First Person Vision Algorithms. Previous works in First Person Vision (FPV) have mainly addressed short-range video understanding tasks based on local temporal context, such as action recognition [29], action anticipation [11], human-object interaction [14, 24], and action detection [34]. More recent stud-

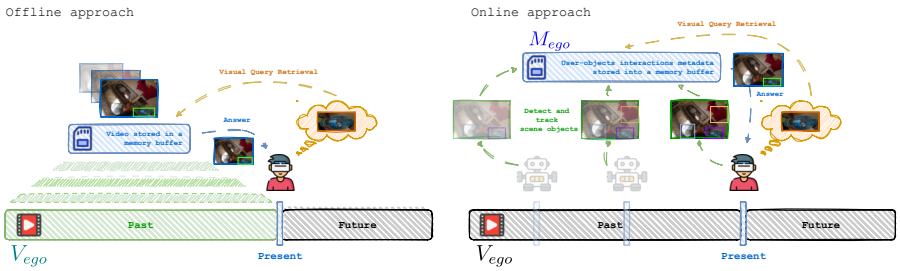

Fig. 1: Offline vs. Online Visual Query Localization. The left diagram illustrates the *offline* approach to visual query localization, where a user queries a previously seen object in the scene. The retrieval system processes the query by accessing the entire video stored in a buffer. In contrast, the right diagram illustrates the proposed *online* approach, which relies on object memorization and retrieval. In this case, the system maintains a memory log of detected objects, removing the need to store the entire video.

ies [7, 12, 14] have explored long-range tasks, such as recalling the position of previously seen objects [7], retrieving video segments with specific objects or events [12], or referencing past scenes to locate target objects [14]. These approaches typically assume an offline setting, where the entire video history is accessible for re-processing. For instance, [12] re-analyze all past video data upon receiving a visual query. Such methods are impractical in streaming scenarios, where the system can only observe the video once and must store limited information for future use.

Visual Query Localization (VQL). The problem of VQL was first formalized by [12] as part of the Episodic Memory (EM) benchmark within the Ego4D dataset. VQL is defined as the task of identifying a spatio-temporal bounding box sequence indicating the last occurrence of an object of interest in a video when provided with a visual patch depicting the object. Early approaches to VQL relied on frame-by-frame matching using architectures such as Siamese R-CNN [27]. A common baseline follows a cascade strategy: it first searches for the query patch in every frame using Siamese R-CNN [27], then applies a short-term tracker [3] forward and backward in time to estimate the temporal extent of the object's presence. Building on this formulation, several methods have explored key aspects to improve performance [6, 13, 15, 20, 31, 32].

The previously mentioned methods are offline approaches, which often face significant storage and computational challenges, making them impractical for real-time applications. In contrast, our method operates online. Rather than storing and processing entire video streams offline, our approach continuously encodes and tracks object instances as the video progresses (Figure 1). This provides a compact memory log, reducing storage requirements, and ensures a memory-efficient system that allows for faster retrieval during user queries.

Object Detection has undergone significant evolution, marked by various methodological developments. Two-stage methods, such as Faster R-CNN [23],

extract feature maps at different scales and use a specialized Region Proposal Network (RPN) to generate object proposals on these feature maps, which are subsequently refined to accurately detect objects. In contrast, the YOLO family [22, 28] adopts a single-stage approach. YOLO predicts bounding boxes and class probabilities simultaneously, thereby increasing speed due to its end-to-end architecture; however, it generally achieves lower detection accuracy compared to other architectures. More recent work, such as DETR [5], leverages transformers and learnable queries to approach object detection through an encoder-decoder structure. Although DETR benefits from self-attention mechanisms, which enhance performance, it lacks the speed necessary for some real-time applications. For our work we used Faster R-CNN due to is optimal trade-off between accuracy and computational efficiency in object detection, making it a default choice in scenarios that demand both high performance and near real-time processing.

Visual Object Tracking (VOT). VOT is a fundamental challenge in computer vision [16, 17], aimed at estimating the trajectory of one – Single-Object Tracking (SOT) – or multiple – Multiple-Object Tracking (MOT) – target objects across video frames. This task is essential for maintaining consistent object identities and ensuring robust performance across frames. For the tracking of a single object, advancements have been made possible thanks to correlation filters [4], siamese neural networks [1, 10], deep discriminative networks [2], and transformers [33]. For tracking multiple objects at the same time, progress has been achieved with tracking-by-detection [30], tracking-by-regression [35], and more recently with tracking-by-attention [19].

The main goal for VOT is to create models that can reliably track a target over time, overcoming challenges such as occlusions, scale variations, deformations, rotations, motion blur, and environmental factors like changing lighting conditions. Among these, occlusion is a particularly difficult issue, which is why VOT is typically divided into two categories: short-term tracking, where the target remains visible throughout the sequence, and long-term tracking, which involves detecting and relocating a target that intermittently disappears and reappears. In the domain of FPV, the challenge intensifies [8, 9]. FPV focuses on processing images captured from wearable cameras—often mounted on the user's head or chest—which provide a perspective distinct from that of traditional fixed cameras. The proximity of the camera to objects, coupled with frequent interactions between the wearer and the environment, introduces additional complexities. Issues such as occlusions, rapid movements, and constant changes in object size and appearance make tracking in these contexts significantly more difficult than in stationary camera setups.

3 Methods

We define the online VQ2D task as follows. Given a streaming video $\mathcal{V} = \{\mathbf{F}_t\}$ of RGB frames $\mathbf{F}_t$ indexed by time t and a query image RGB patch $\mathbf{Q}$ representing the visual appearance of the query object, the goal is to retrieve a response track $\mathbf{r} = \{(\mathbf{F}_s, \mathbf{b}_s),(\mathbf{F}_{s+1}, \mathbf{b}_{s+1}), ...,(\mathbf{F}_e, \mathbf{b}_e)\}$, which is an ordered sequence of

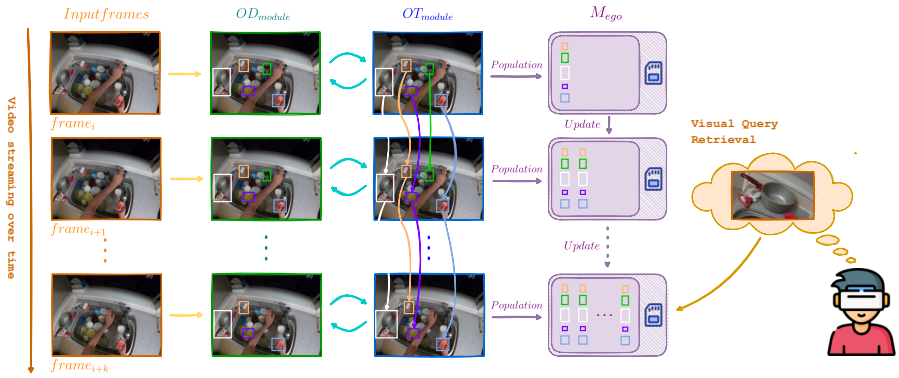

Fig. 2: T-EVO framework. Overview of the proposed framework consisting of three main modules: an Object Detection module (OD_{module}) that identifies and localizes objects; an Object Tracking module (OT_{module}) that associates object instances across frames; and a Memory module ($\mathcal{M}_{ego}$) that stores and updates object representations for future retrieval. This architecture enables users to perform visual query-based object retrieval efficiently.

pairs $(\mathbf{F}_j, \mathbf{b}_j)$ of frame and bounding box, where $s \leq j \leq e \leq t$. This track determines the spatio-temporal position of the object's last appearance in $\mathcal{V}$, with s and e denoting starting and ending frames, and each box $\mathbf{b}_j$ corresponding to frame $\mathbf{F}_j$. The goal is to find $\mathbf{r}$ given $\mathbf{Q}$, processing $\mathcal{V}$ in streaming fashion, observing frames $\mathbf{F}_t$ only once. To support future queries, we store information in a compact memory.

3.1 Proposed Framework

Inspired by the human ability to understand long-range user-object interactions, we propose to study an algorithmic pipeline with the following objectives: 1) Discover Objects (OD): identify significant objects to be tracked and stored in the memory; 2) Object Tracking (OT): track the discovered objects in an instance-based manner over extended periods; 3) High-Level compact Memory ($\mathcal{M}_{ego}$): utilize information about the detected objects and their associated tracks to create high-level and compact memories that describe past user interactions; 4) Memory Retrieval (MR): retrieves the visual response track $\mathbf{r}$ for a visual query $\mathbf{Q}$ from memory. Figure 2 illustrates the structure of the proposed framework.

OD module. The object discovery module is responsible for determining which objects are potentially useful to track and include in the high-level memories to be queried by the user in a later stage. The object discovery module is implemented as an object detector that detects a number B of bounding boxes $\mathbf{b}_{k,t}$ in the frame $\mathbf{F}_t$, coupled with a confidence score $s_{k,t}$, where $1 \leq k \leq B$. The detected objects are first filtered based on the confidence scores $s_{k,t} > \lambda_{od}$. The bounding box $\mathbf{b}_{i,t}$ related to the detected object are compared with those already present in the memory at time t. These are objects that might have

been already tracked at time t from object instances at time $t-1$. If $s_{k,t} > \lambda_{od}$ and $IoU > \lambda_{iou}$ with any of the objects present in the memory, the object is considered as a new instance to include in the memory at time t.

OT module. The OT module tracks the detected objects in a long-term fashion. The module aims to update the object localization, keeping the same ID. Temporally continuous information is fundamental to know where specific object instances move or how their state changes. At each time t, the module retrieves the location of the objects $\mathbf{O}_i$ stored in the memory and then estimates the spatial position $\mathbf{b}_{i,t}$ in the frame $\mathbf{F}_t$ by means of a visual tracker. For each $\mathbf{b}_{i,t}$ the tracker outputs also a confidence score $s_{i,t}$. Only the $\mathbf{b}_{i,t}$ with $s_{i,t} > \lambda_{ot}$ with the associated frame $\mathbf{F}_t$ are retained in the memory.

Memory ($\mathcal{M}_{ego}$): The third component is the memory. We define the object memory as a set $\mathcal{M}_{ego}= \{\mathbf{O}_1, \ldots, \mathbf{O}_i, \ldots, \mathbf{O}_N\}$ of representations $\mathbf{O}_i$ of the spatio-temporal locations of objects in the frames $\mathbf{F}_t$ of $\mathcal{V}$. The indexes $1 \leq i \leq N$ denote object instance unique IDs. In practice, $\mathbf{O}_i$ is a list $\mathbf{O}_i = \{\mathbf{o}_{i,t}\}$ of records temporally indexed by t, with each record having the form $\mathbf{o}_{i,t} = [t, \mathbf{b}_{i,t}, \mathbf{F}_t]$, where $\mathbf{b}_{i,t} = [\mathbf{b}_{x_{i,t}}, \mathbf{b}_{y_{i,t}}, \mathbf{b}_{w_{i,t}}, \mathbf{b}_{h_{i,t}}]$ denotes the object bounding box, $\mathbf{F}_t$ is the associated RGB frame at time t. If $\mathbf{o}_{i,t}$ is not visible in $\mathbf{F}_t$, then $\mathbf{o}_{i,t} = \emptyset$. This formulation provides an indexed, lightweight, compact, human-readable, and easily queryable representation of objects observed in $\mathcal{V}$ up to the current frame $\mathbf{F}_t$.

Memory Retrieval. The Memory Retrieval module is responsible for retrieving the visual response track $\mathbf{r}$ of a visual query $\mathbf{Q}$ within the memory. The module takes as an input the visual query $\mathbf{Q}$ and outputs a sequence of contiguous bounding boxes $\mathbf{b}_{i,t}$ and the associated frames $\mathbf{F}_t$ for the object instance $\mathbf{O}_i$ that matches $\mathbf{Q}$. The procedure starts by extracting the query's feature representation, denoted as $\phi(\mathbf{Q})$, using a feature extraction function $\phi(.)$. The features $\Phi(\phi_{i,t})$ of the object representations $\phi_{i,t}$ are then compared with the features of the visual query $\mathbf{Q}$, using a similarity function $\Psi(\phi(\mathbf{Q}), \Phi(\phi_{i,t}))$ that outputs values in the range [0,1]. The similarity scores obtained for multiple bounding boxes $\mathbf{b}_{i,t}$ of the same object i are averaged into a single score $\mathbf{s}_i$. If the highest score $\mathbf{s}_i$ among all objects exceeds the threshold λ_{ret}, the object $\mathbf{O}_i$ is identified as a match for the query $\mathbf{Q}$. Consequently, the most recent sequence of contiguous bounding boxes and frames from $\mathbf{O}_i$'s spatio-temporal history is returned as the visual response track $\mathbf{r}$.

4 Experimental settings

Implementation Details. For the Object Discovery (OD) module, we employed a Faster R-CNN R101-FPN detector ($OD_{fasterrcnn}$) [23] fine-tuned on EgoTracks [25]. Detections have been filtered with a confidence score threshold $\lambda_{od} = 0.05$. To evaluate optimal detections performances, we also considered ground-truth EgoTracks annotations (OD_{oracle}). EgoTracks aligns with the Episodic Memory benchmark of Ego4D, ensuring that its detections contain valid queries.

For the Object Tracking Module (OT) module, we employed EgoSTARK [25],

a STARK [33] instance fine-tuned on EgoTracks ($OT_{egostark}$). $OT_{egostark}$ implements a multiple-object tracking approach using multiple instances of single-object trackers. This approach is motivated by the lack of effective multi-object tracking solutions for egocentric vision. To track multiple objects, we deployed multiple EgoSTARK trackers simultaneously. A new tracker is initialized in frame F_t whenever the OD module detects a valid object, which is then stored in memory. Each bounding box predicted by the OD module acts as a reference to initialize a corresponding tracker. Once initialized, the tracker follows the object across subsequent frames, updating its spatio-temporal representation in the memory. To evaluate optimal tracking performances, we also considered ground-truth EgoTracks annotations (OT_{oracle}).

In memory, for each object $\mathbf{O}_i$, the bounding boxes $b_{i,t}$ and the associated frames F_t from the most recent response track are stored if the confidence score $s_{i,t} > \lambda_{ot} = 0.5$.

To localize the visual query $\mathbf{Q}$ in memory, we implemented an approach based on SiamRCNN with a classification head [27]. The SiamRCNN approach, utilizes a Feature Pyramid Network (FPN) [18] as its backbone $\phi(.)$. This network generates feature representations for both the visual memory of objects and the visual query $\mathbf{Q}$. To assess whether $\Phi(\mathbf{Q})$ is stored in memory, the feature $\Phi(\phi_{i,t})$ of each object is compared to $\Phi(\mathbf{Q})$ using a Siamese network head [27]. This head performs a bilinear operation that predicts an instance similarity score within the interval [0, 1]. The similarity scores are aggregated into a single value $\mathbf{r_i}$. If the highest score $\mathbf{r_i}$ surpasses the threshold λ_{ref}, the object $\mathbf{O_i}$ is deemed a match for the query $\mathbf{Q}$. Upon identifying a match, the latest sequence of consecutive bounding boxes from the spatio-temporal history of $\mathbf{O_i}$ is retrieved and provided as the response track $\mathbf{r}$.

Dataset. We relied on publicly available datasets to conduct our evaluations. Specifically, we evaluated our framework on Ego4D Episodic Memory VQ2D benchmark [12], the only publicly available dataset for Visual Query Localization. This benchmark includes 433 hours from 54 scenarios, featuring 22K visual queries spanning 3K object categories. For evaluation, we used 115 video clips from the validation split of the Ego4D VQ2D benchmark. They contain 450 visual queries along with the corresponding ground truth response track and tracklets. The total amount of frames is 208,365, and we run the algorithm at the original frame rate of 5 frames per second. To train our components, we used EgoTracks [25], an egocentric video dataset sourced from Ego4D and aligned to the episodic memory benchmark. EgoTracks provides sparse object annotations across contiguous frames. To train the detector on the EgoTracks benchmark, we conducted a clustering-based refinement of the object taxonomy. Since object instances are associated with textual annotations, we initially applied K-Means clustering to group object descriptions, yielding 1,169 preliminary classes. We then manually refined these clusters to improve accuracy and consistency, resulting in a diverse and well-structured taxonomy comprising 295 classes.

Evaluation Protocol and Metrics. Each video clip $\mathcal{V}$ is processed to construct the object memory $\mathcal{M}_{ego}$. The memory retrieval algorithm then analyzes each

Table 1: Comparison of the performance of different T-EVO configurations for the Online Visual Query 2D task, compared to offline approaches. Gray rows display oracular results. Performance for the online task shows a significant decrease when compared to offline results. However, a substantial reduction in memory usage and retrieval time is observed in the online configuration.

Method	OD	OT	%tAP$_{25}$ ↑	%stAP$_{25}$ ↑	%Succ ↑	Size ↓	Time ↓
Offline Methods (VQ2D)							
SiamRCNN+KYS	/	/	20.0	12.0	39.8	12.1 GB	8.3 m
STARK	/	/	10.0	4.0	18.7	12.1 GB	45 s
SiamRCNN	/	/	22.0	15.0	43.2	12.1 GB	8.3 m
CocoFormer	/	/	27.0	20.0	48.4	12.1 GB	8.3 m
VQLoC	/	/	31.0	22.0	55.9	12.1 GB	41 s
Online Methods (OVQ2D)							
T-EVO	OD$_{\text{oracle}}$	OT$_{\text{oracle}}$	73.3	68.3	81.9	504.7 MB	4.3 s
		OT$_{\text{egostark}}$	5.9	3.7	25.9	230.1 MB	1.94 s
	OD$_{\text{fasterrcnn}}$	OT$_{\text{oracle}}$	19.6	11.2	36.1	190.2 MB	1.60 s
		OT$_{\text{egostark}}$	0.1	0.1	2.9	2.9 GB	24.93 s

visual query We evaluate performance for downstream the online VQ2D task using standard VQL metrics [12]: %tAP$_{25}$ ↑measures temporal alignment with the ground truth at a 0.25 IoU threshold; %stAP$_{25}$ ↑assesses spatio-temporal precision at a 0.25 spatio-temporal IoU threshold; %Succ ↑calculates the proportion of predictions with at least 0.05 spatio-temporal IoU. All metrics are reported within the range [0, 100].
Additionally, we measure efficiency in terms of storage size (Size ↓, in MB/GB) required for query localization and retrieval time (Time ↓, in seconds/minutes) needed to localize a query.

5 Results

Online VQ2D Benchmark. Table 1 presents the performance of various T-EVO configurations for the online VQ2D task. This task proves particularly challenging, with a success rate of only 2.9%. Oracular results show that combining OT$_{egostark}$ with oracle detections improves success (36.1), OD$_{fasterrcnn}$ with an oracular tracker reaches 25.9, and full oracular detection and tracking achieve 81.9. These findings highlight T-EVO 's potential while underscoring the need for better detection and tracking in real-world egocentric video. Unlike prior studies evaluating these components in isolation, this benchmark provides a principled testbed for assessing their real-world performance in downstream tasks.
Comparison with Offline VQ2D. Table 1 compares offline VQ2D methods with T-EVO. While not directly comparable, online VQ2D is a harder task since the methods lack prior knowledge of the queried object. These results mainly highlight the differences between the two approaches.
Offline methods require storing entire video sequences, leading to high memory

Table 2: Comparison of T-EVO's performance at different video progress levels. We evaluate retrieval accuracy, storage usage, and retrieval time at 25%, 50%, 75%, and 100% of the video for all OD and OT T-EVO configurations. Results show that the best performance is achieved at 50% with $\text{OT}_{egostark}$, indicating that long-term tracking degrades as the video progresses.

OD	OT	Percentage	%Succ ↑	Size ↓	Time ↓
$\text{OD}_{\text{oracle}}$	$\text{OT}_{\text{oracle}}$	25%	75.3	**62.0 MB**	**0.5 s**
		50%	82.6	100.9 MB	0.9 s
		75%	**82.8**	276.6 MB	2.3 s
		100%	81.9	504.7 MB	4.3 s
	$\text{OT}_{\text{egostark}}$	25%	23.5	**215.4 MB**	**1.8 s**
		50%	**32.4**	220.4 MB	1.9 s
		75%	28.7	295.8 MB	2.5 s
		100%	25.9	230.1 MB	1.9 s
$\text{OD}_{\text{fasterrcnn}}$	$\text{OT}_{\text{oracle}}$	25%	17.7	**124.3 MB**	**1.1 s**
		50%	36.1	184.4 MB	1.6 s
		75%	**38.7**	242.6 MB	2.0 s
		100%	36.0	190.2 MB	1.6 s
	$\text{OT}_{\text{egostark}}$	25%	0.1	**732.1 MB**	**6.6 s**
		50%	**8.1**	1.5 GB	12.2 s
		75%	4.0	2.2 GB	18.8 s
		100%	2.9	2.9 GB	24.9 s

usage (e.g., 12.1 GB for a 5-minute clip). T-EVO significantly reduces memory (up to 4× smaller, 2.9 vs. 12.1 GB) and speeds up inference (2× faster, 24.9 s vs. 41 s compared to VQLoC). In oracular settings, T-EVO achieves even greater efficiency (24× less memory, 9× faster) while also being competitive in success rate (81.9 vs. 55.9 for VQLoC), demonstrating its potential when detection and tracking are improved.

Retrieval at different portion of the video. Table 2 reports retrieval performance, storage requirements, and retrieval time as the video progresses through 25%, 50%, 75%, and 100% of its length, across all OD and OT T-EVO configurations. Performance is lowest at 25%, likely because many query-relevant objects have not yet been observed. The highest performance is achieved at 50% with $\text{OT}_{egostark}$, suggesting that tracking quality degrades over time, affecting long-term retrieval. In contrast, with OT_{oracle}, the best performance occurs at 75%, likely because most query-relevant objects have by then been stored in memory.

Qualitative Results. Figure 3 shows a qualitative example of memory retrieval, localizing a visual query within two T-EVO's configurations.

6 Conclusions

In this paper, we introduced a preliminary framework for addressing the Visual Query 2D (VQ2D) task in an online setting. The proposed framework, named T-EVO, includes an object detection module and an object tracking module, which

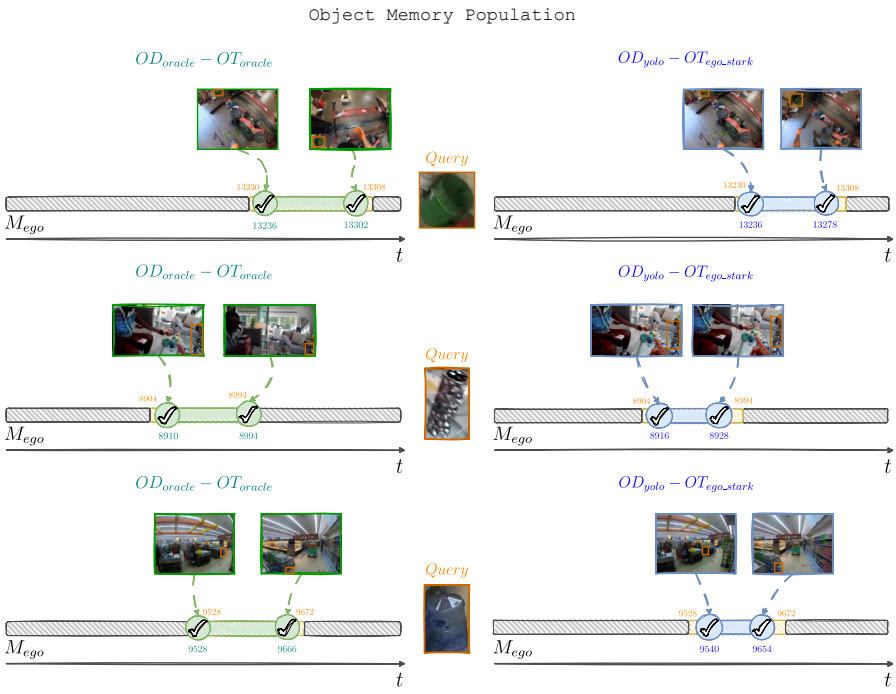

Fig. 3: Qualitative results. The results of memory retrieval from $\mathcal{M}_{ego}$ are presented using OD_{oracle}-OT_{oracle} and $\text{OD}_{fasterrcnn}$-$\text{OT}_{egostark}$. The figures display the query **Q** along with the corresponding frames and bounding boxes belonging to the response track stored in the memory ($\mathcal{M}_{ego}$).

together populate a memory to enable efficient query retrieval. Oracle-based results showcase the framework's effectiveness, achieving a success rate of 81.9%. However, when replacing the oracle components with real-world detectors and trackers, performance drops significantly to 2.9%, highlighting the limitations of current systems. Despite this, T-EVO maintains a compact memory structure, reducing storage requirements by a factor of 24 and accelerating retrieval by a factor of 9, demonstrating its strong potential for deployment in real-world wearable devices.

Acknowledgments. Progetto PRIN 2022 PNRR - "Tracking in Egovision for Applied Memory (TEAM)" Codice P20225MSER_001. Codici CUP G53D23006680001 (Università di Udine) e E53D23016240001 (Università di Catania). This research has been funded by the European Union, NextGenerationEU – PNRR M4 C2 I1.1 RS Micheloni.

References

1. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCVW (2016)
2. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: ICCV (2019)
3. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Know your surroundings: Exploiting scene information for object tracking. In: ECCV (2020)
4. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: CVPR (2010)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
6. Chen, G., Xing, S., Chen, Z., Wang, Y., Li, K., Li, Y., Liu, Y., Wang, J., Zheng, Y.D., Huang, B., et al.: Internvideo-ego4d: A pack of champion solutions to ego4d challenges. arXiv preprint arXiv:2211.09529 (2022)
7. Datta, S., Dharur, S., Cartillier, V., Desai, R., Khanna, M., Batra, D., Parikh, D.: Episodic memory question answering. In: CVPR (2022)
8. Dunnhofer, M., Furnari, A., Farinella, G., Micheloni, C.: Is first person vision challenging for object tracking? In: ICCVW (2021)
9. Dunnhofer, M., Furnari, A., Farinella, G.M., Micheloni, C.: Visual object tracking in first person vision. IJCV (2023)
10. Dunnhofer, M., Martinel, N., Micheloni, C.: Weakly-supervised domain adaptation of deep regression trackers via reinforced knowledge distillation. IEEE Robotics and Automation Letters (2021)
11. Furnari, A., Farinella, G.M.: Rolling-unrolling lstms for action anticipation from first-person video. IEEE TPAMI (2020)
12. 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)
13. Jiang, H., Ramakrishnan, S.K., Grauman, K.: Single-stage visual query localization in egocentric videos. NeurIPS (2023)
14. Khalil, A.A.D., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fouhey, D., Fidler, S., Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations. In: NeurIPS (2022)
15. Khosla, S., TV, S., Schwing, A., Hoiem, D.: Relocate: A simple training-free baseline for visual query localization using region-based representations. arXiv preprint arXiv:2412.01826 (2024)
16. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.K., Chang, H.J., Danelljan, M., Zajc, L.Č., Lukežič, A., et al.: The tenth visual object tracking vot2022 challenge results. In: ECCVW (2022)
17. Kristan, M., Matas, J., Danelljan, M., Felsberg, M., Chang, H.J., Zajc, L.Č., Lukežič, A., Drbohlav, O., Zhang, Z., Tran, K.T., et al.: The first visual object tracking segmentation vots2023 challenge results. In: ICCVW (2023)
18. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
19. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: CVPR (2022)
20. Pei, B., Chen, G., Xu, J., He, Y., Liu, Y., Pan, K., Huang, Y., Wang, Y., Lu, T., Wang, L., et al.: Egovideo: Exploring egocentric foundation model and downstream adaptation. arXiv preprint arXiv:2406.18070 (2024)

21. Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G.M., Damen, D., Tommasi, T.: An outlook into the future of egocentric vision. IJCV (2024)
22. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE TPAMI (2016)
24. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: CVPR (2020)
25. Tang, H., Liang, K.J., Grauman, K., Feiszli, M., Wang, W.: Egotracks: A long-term egocentric visual object tracking dataset. NeurIPS (2023)
26. Tulving, E.: Episodic memory: From mind to brain. Annual review of Psychology (2002)
27. Voigtlaender, P., Luiten, J., Torr, P.H., Leibe, B.: Siam r-cnn: Visual tracking by re-detection. In: CVPR (2020)
28. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. NeurIPS (2024)
29. Wang, X., Wu, Y., Zhu, L., Yang, Y.: Symbiotic attention with privileged information for egocentric action recognition. In: AAAI (2020)
30. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP (2017)
31. Xu, M., Fu, C.Y., Li, Y., Ghanem, B., Perez-Rua, J.M., Xiang, T.: Negative frames matter in egocentric visual query 2d localization. arXiv preprint arXiv:2208.01949 (2022)
32. 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)
33. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: ICCV (2021)
34. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: ECCV (2022)
35. Zhou, P., Chi, M.: Relation parsing neural network for human-object interaction detection. In: ICCV (2019)