EAGLE: Episodic Appearance- and Geometry-aware Memory for Unified 2D-3D Visual Query Localization in Egocentric Vision

Anonymous submission

Abstract

Egocentric visual query localization is vital for embodied AI and VR/AR, yet remains challenging due to camera motion, viewpoint changes, and appearance variations. We present **EAGLE**, a novel framework that leverages **e**pisodic **a**ppearance- and **g**eometry-aware memory to achieve unified 2D-3D visual query **l**ocalization in **e**gocentric vision. Inspired by avian memory consolidation, EAGLE synergistically integrates segmentation guided by an appearance-aware meta-learning memory (AMM), with tracking driven by a geometry-aware localization memory (GLM). This memory consolidation mechanism, through structured appearance and geometry memory banks, stores high-confidence retrieval samples, effectively supporting both long- and short-term modeling of target appearance variations. This enables precise contour delineation with robust spatial discrimination, leading to significantly improved retrieval accuracy. Furthermore, by integrating the VQL-2D output with a visual geometry grounded Transformer (VGGT), we achieve a efficient unification of 2D and 3D tasks, enabling rapid and accurate back-projection into 3D space. Our method achieves state-of-the-art performance on the Ego4D-VQ benchmark. Code will be released soon.

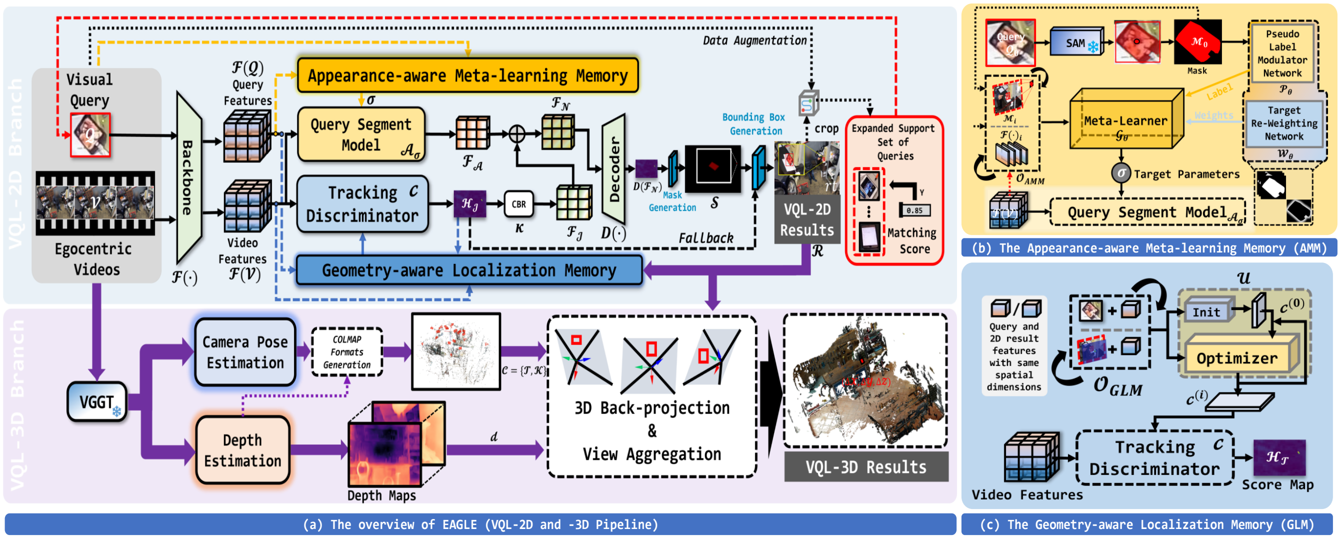
1 Introduction

Visual query localization (VQL) is a fundamental task in egocentric episodic memory, which aims to spatio-temporally localize the final occurrence of a target within a video, guided by a visual crop. This capability serves as a cornerstone for downstream applications such as virtuality or augmented reality (VR/AR), embodied AI, etc..(Grauman et al. 2022; Plizzari et al. 2024; Jiang, Ramakrishnan, and Grauman 2024; Mai et al. 2023; Xu et al. 2023; Tang et al. 2024; Hao et al.) Nevertheless, the egocentric perspective presents challenges, including drastic camera motion, severe motion blur, and variations in object appearance and scale. These factors frequently lead to retrieval failures, critically impeding progress in this field.

We revisit the prevalent ”detect-then-track” pipeline for VQL and expose its limitations. (i) The aforementioned paradigm couples a detector (the *identifier*) with a tracker (the *navigator*), manifests following principal shortcomings: the identifier’s reliance on bounding boxes leads to the inclusion of substantial background pixels, especially for non-rigid targets. These view-dependent background signals can contaminate the target’s appearance, thus impairing search accuracy. Concurrently, the navigator lacks the robustness to handle challenges such as extreme view changes, drastic scale transformations, motion blur, and similar distractors. (ii) During retrieval, relying on a static, low-visibility query is often insufficient to capture the target’s appearance variations over time; instead, humans typically integrate multiple visual cues from different temporal snapshots to compensate for the limitations of a single-frame query. (iii) The prevailing VQL paradigms have not yet achieved a natural unification between 2D and 3D tasks, notwithstanding the intimate correlation between them in real world.

The avian visual system offers profound insights into this problem. Eagle, for instance, exhibit remarkable episodic memory and spatial localization, enabling them to retain the appearance of a specific object over extended periods and precisely recall its spatiotemporal position(Bevandić et al. 2024; Dickerson and Eichenbaum 2010). This capability originates from a ”memory consolidation” mechanism: initially, the system rapidly forms a short-term memory ”imprint” of key features. Subsequently, through continuous observation, it actively disambiguates the target from its evolving environment, eventually solidifying this information into a stable, long-term memory. Inspired by this biological schema, an ideal VQL system should transcend the passive storage of static visual cues. It must instead actively filter and encode object that is both critical for long-term recognition and inherently stable. Such a selective memorization mechanism enhances the model’s robustness to environmental distractors, thereby facilitating precise target retrieval within dynamic and variable egocentric videos.

To address the aforementioned challenges, we introduce a novel VQL framework that integrates two parallel branches—segmentation and tracking—which serve as the *identifier* and *navigator*, respectively. Inspired by the memory consolidation mechanism in avian vision systems, both branches are driven by independent online episodic memory banks. These banks dynamically update their support sets by continuously filtering for high-confidence visual observations, enabling long-term adaptive learning of the target’s state. Specifically, the segmentation branch incorporates an appearance-aware meta-learning memory (AMM). Starting with an initial mask generated by SAM(Kirillov et al. 2023), it constructs rich supervisory signals via a pseudo-


VQL-2D Branch
Visual Query
$\mathcal{Q}$
$\mathcal{V}$
Egocentric Videos
Backbone
$\mathcal{F}(\cdot)$
$\mathcal{F}(\mathcal{Q})$ Query Features
Video Features $\mathcal{F}(\mathcal{V})$
Appearance-aware Meta-learning Memory
σ
Query Segment Model $\mathcal{A}_\sigma$
$\mathcal{F}_\mathcal{A}$
$\mathcal{F}_\mathcal{N}$
Tracking $\mathcal{C}$ Discriminator
$\mathcal{H}_\mathcal{J}$
CBR
κ
$\mathcal{F}_\mathcal{J}$
Decoder
$D(\cdot)$
$D(\mathcal{F}_\mathcal{N})$
Mask Generation
$\mathcal{S}$
Bounding Box Generation
Fallback
Data Augmentation
crop
VQL-2D Results
$\mathcal{R}$
Expanded Support Set of Queries
Y
0.85
Matching Score
Geometry-aware Localization Memory
VQL-3D Branch
VGGT
Camera Pose Estimation
COLMAP Formats Generation
$\mathcal{C} = \{\mathcal{T}, \mathcal{K}\}$
Depth Estimation
Depth Maps
d
3D Back-projection & View Aggregation
$(\Delta x, \Delta y, \Delta z)$
VQL-3D Results
Query $\mathcal{Q}_0$
SAM
Mask
$\mathcal{M}_0$
Pseudo Label Modulator Network
$\mathcal{P}_\theta$
Label
Weights
Target Re-Weighting Network
$\mathcal{W}_\theta$
$\mathcal{M}_i$
$\mathcal{F}(\cdot)_i$
Meta-Learner $\mathcal{G}_\theta$
$\mathcal{O}_{AMM}$
Target Parameters
$\mathcal{F}(\mathcal{V})$
Query Segment Model $\mathcal{A}_\sigma$
Query and 2D result features with same spatial dimensions
$\mathcal{O}_{GLM}$
$\mathcal{U}$
Init
$c^{(0)}$
Optimizer
$c^{(i)}$
Tracking $\mathcal{C}$ Discriminator
Video Features
$\mathcal{H}_\mathcal{T}$
Score Map

(a) The overview of EAGLE (VQL-2D and -3D Pipeline)

(b) The Appearance-aware Meta-learning Memory (AMM)

(c) The Geometry-aware Localization Memory (GLM)

Figure 1: **Overview of EAGLE**. Our framework consists of two branches: VQL-2D and -3D. The VQL-2D features a dual-branch architecture built upon a shared feature backbone. The segmentation branch serves as a precise identifier, guided by an appearance-aware meta-learning memory (b) to generate pixel-level masks with fine-grained semantic cues. The tracking branch, acting as a navigator, is driven by a geometry-aware localization memory (c) to produce a discriminative score map robust to egocentric view changes. Finally, a decoder fuses the outputs from both branches to yield the final results. VQL-3D paradigm leverages the VGGT to jointly process the 2D results, camera pose, and depth, ultimately predicting the positional offset of the target in 3D space.

label modulator network and a target re-weighting network. This module populates its episodic memory bank with high-confidence retrieval results and employs the steepest descent method to rapidly update model parameters, facilitating continuous meta-learning of the target's appearance for pixel-level accurate retrieval. The tracking branch employs a discriminative correlation filter (DCF)(Bolme et al. 2010) to rectify potential instance misassignments from the segmentation branch. By maintaining a memory bank that stores both the static initial query and dynamic high-confidence observations, this branch proactively encodes stable geometric information crucial for long-term recognition. This provides strong constraints on the segmentation results and enables robust, long-term modeling of target appearance and scale variations. Furthermore, to efficiently unify 2D and 3D tasks, we feed the 2D-VQL output into a VGGT(Wang et al. 2025) to refine camera pose and depth estimations. Subsequently, the 2D segmentation results are back-projected into 3D space, yielding a unified 3D-VQL output. This design significantly enhances inference efficiency and memory utilization, laying a solid foundation for more realistic embodied episodic memory retrieval. In summary, the contributions are as follow:

- We propose a novel dual-branch framework for egocentric VQL that synergizes segmentation and tracking to leverage precise, pixel-level appearance cues and robust geometric-temporal constraints, overcoming the limitations of traditional "detect-then-track" paradigms.

- Inspired by biological cognitive processes, we devise an online episodic memory bank-driven memory consolidation mechanism for both branches. This mechanism selectively incorporates high-confidence observations into structured appearance and geometry memory banks, enabling continuous, long-term adaptation to target variations while mitigating interference from distractors..

- We achieve an efficient and unified 2D-to-3D localization by back-projecting the refined 2D VQL outputs into 3D space, significantly enhancing the applicability of our method for embodied AI scenarios.

- Comprehensive experiments on the Ego4D-VQ benchmark validate the superiority of our method, demonstrating state-of-the-art performance in accuracy, robustness, and efficiency.

2 Related Work

Few-shot Visual Object Tracking. VQL is fundamentally a few-shot, open-world tracking problem(Xu et al. 2022; Fan et al. 2025; Khosla et al. 2025; Tang et al. 2024; Zhao et al. 2024). While meta-learning-based trackers(Lukezic, Matas, and Kristan 2020; Bhat et al. 2019; Choi, Kwon, and Lee 2019; Park and Berg 2018; Dai et al. 2020; Bhat et al. 2020; Yan et al. 2019) show promise due to their rapid adaptation capabilities, they primarily address "how to update." We argue that the core challenge in VQL is "what to learn"—constructing an optimal and robust target representation. To this end, inspired by biological memory consolidation, we introduce a dual-memory mechanism. Our model jointly leverages segmentation and tracking to dynamically build online memory banks from query-response interactions within the video, enabling it to actively distill discriminative features and robustly handle the target's evolving states.

2D & 3D Visual Query Localization. VQL, comprising both 2D and 3D tasks, aims to spatio-temporally localize a target's final occurrence. Early methods for VQL-2D relied on multi-stage "detect-then-track" pipelines (Grauman et al. 2022; Xu et al. 2023, 2022), while more recent works like PRVQL (Fan et al. 2025) and RELOCATE (Khosla et al. 2025) have streamlined this into single-stage or training-free frameworks. For VQL-3D, existing approaches (Forigua et al. 2023; Mai et al. 2023, 2024) typically depend on Structure-from-Motion (SfM)(Ullman 1979) techniques like COLMAP(Schonberger and Frahm 2016) for pose estimation. However, SfM is often fragile, failing in texture-less or high-motion scenarios, and leads to fragmented, inefficient pipelines. To overcome these limitations, we first address VQL-2D by integrating memory-guided segmentation and tracking for robust retrieval. We then tackle VQL-3D by replacing the brittle SfM-based pose estimation with VGGT (Wang et al. 2025). Compared to other learning-based geometric foundation models(Mildenhall et al. 2021; Tschernezki, Larlus, and Vedaldi 2021; Kerbl et al. 2023; Wang et al. 2024), VGGT more efficiently infers camera pose and depth in a feedforward pass. This substitution not only enhances 3D displacement accuracy but also unifies the 2D and 3D tasks into a single, efficient pipeline.

3 The Proposed Method

Figure.1 (a) illustrates the framework of EAGLE. Our approach is detailed in the following sections, and the preliminaries for VQL is provided in Appendix A.1.

3.1 The AMM-guided segmentation branch

As shown in Figure.1(b), we leverage SAM to obtain an initial segmentation mask, denoted as $\mathcal{M}_0 = SAM(\mathcal{Q}_0)$, for the visual query $\mathcal{Q}_0$. To provide richer supervision for the subsequent meta-learning process, we introduce a pseudo-label modulator network, $\mathcal{P}_\theta$. This trainable, lightweight convolutional network transforms the input binary mask into a multi-channel pseudo-label, $\mathcal{QM}_i = \mathcal{P}_\theta(\mathcal{M}_i)$, which encapsulates rich semantic information such as boundaries and centers. Concurrently, we design a target re-weighting network ($\mathcal{W}_\theta$) with a similar architecture to $\mathcal{P}_\theta$, aiming to guide the loss function to focus on critical regions of the target. To facilitate an efficient and differentiable meta-learning process, $\mathcal{A}_\sigma$ maps a tensor from $\mathbb{R}^{H\times W\times C}$ to $\mathbb{R}^{H\times W\times D}$. This is formulated as $\mathcal{A}_\sigma(x) = x * \sigma$, where σ represents the weights of a $\mathbb{R}^{K\times K\times C\times D}$ convolutional layer, and $*$ denotes the convolution operation. Our proposed meta-learner, $\mathcal{G}_\theta$, optimizes the parameters σ by minimizing the weighted squared error between the prediction of the query segment model $\mathcal{A}_\sigma$ and the corresponding pseudo-labels over the episodic memory bank $\mathcal{O}_{AMM}$. Initially, $\mathcal{O}_{AMM}$ contains only the query sample $(\mathcal{F}_{\mathcal{Q}_0}, \mathcal{M}_0)$. Subsequently, it is updated by incorporating segmentation retrieval results extracted from the video stream. Specifically, a segmentation sample based on the segmentation result $\mathcal{S}_{I_i}$ generated by the model on a retrieved video frame I_i is added to the memory bank only if the target exhibits low entropy or high certainty in its corresponding confidence map. This condition is met when the mean confidence score within the predicted mask region exceeds a threshold, $s_{conf} >= 0.6$ [1]. The corresponding region's segmentation sample is then cropped and resized proportionally ($\mathcal{M}_i = \text{crop}(\mathcal{S}_{I_i})$) before being added to the memory bank [2], updating it as $\mathcal{O}_{AMM} = \mathcal{O}_{AMM} \cup \{(\mathcal{F}(I_i), \mathcal{M}_i)\}$. The loss function is then defined as:

$$\mathcal{L}(\sigma) = \frac{1}{2}\sum ||\mathcal{W}_\theta(\mathcal{M}_i) \odot (\mathcal{A}_\sigma(\mathcal{F}_i) - \mathcal{QM}_i)||^2 + \frac{\delta}{2}||\sigma||^2, \tag{1}$$

where δ is a learnable regularizer. $(\mathcal{F}_i, \mathcal{M}_i)$ is the i-th feature-pseudo-label pair from the memory $\mathcal{O}_{AMM}$. Given that Eq.(1) defines a convex quadratic objective with respect to σ, admitting a closed-form solution in either its primal or dual form, we solve it using the iterative steepest descent method. At each iteration, given the current estimate σ_i, the step size α^i is chosen to minimize the loss along the gradient direction, i.e., $\alpha^i = \arg\min_\alpha \mathcal{L}(\sigma_i - \alpha g^i)$, where $g^i = \nabla\mathcal{L}(\sigma_i)$ represents the gradient of the loss function evaluated at σ_i. The parameters are iteratively updated using the steepest descent method [3], as below:

$$\begin{aligned} \sigma_{i+1} &= \sigma_i - \alpha^i g^i, \\ \alpha^i &= \frac{\|g^i\|^2}{\sum_i \|\mathcal{W}_\theta(\mathcal{M}_i) \odot (\mathcal{F}_i * g^i)\|^2 + \delta\|g^i\|^2}, \\ g^i &= \sum_i \mathcal{F}_i *^T \left(\mathcal{W}_\theta^2(\mathcal{M}_i) \odot (\mathcal{F}_i * \sigma_i - \mathcal{P}_\theta(\mathcal{M}_i))\right) + \delta\sigma_i. \end{aligned} \tag{2}$$

where $*^T$ represents the transposed convolution. After i iterations, the resulting parameter σ_i is differentiable with respect to all network parameters θ. The function $\mathcal{G}_\theta(\mathcal{O}_{AMM}, \sigma_0) = \sigma_N$ is defined by performing N iterations of the steepest descent update in Eq.(2), initialized with σ_0. Leveraging the rapid convergence of the steepest descent method, our optimization-based paradigm facilitates efficient updates to σ. New samples are added to $\mathcal{O}_{AMM}$, and a few iterations are performed starting from the current estimate of σ (Eq.(2)).

3.2 The GLM-guided tracking branch

To mitigate potential instance misassignments from the segmentation branch, we introduce a tracking-based navigator. This branch employs DCF to provide robust geometric constraints on the segmentation output. We construct a geometric localization memory (GLM) bank, $\mathcal{O}_{GLM}$, which stores both the initial static query and a dynamic set of high-confidence historical observations retrieved by the 2D branch. Figure.1(c) illustrates the architecture of the GLM. The initial query features and their corresponding Gaussian labels are stored as static snapshots in the episodic memory bank, $\mathcal{O}_{GLM}$. Concurrently, feature maps corresponding to the 2D retrieved target regions from the query-expanded support set are isotropically scaled by a factor of 1.5. These scaled features, along with their Gaussian labels, are then

[1] Design and ablation details in Sec.3.4 and Appendix A.6, respectively

[2] Appendix A.6 details the crop ratio settings.

[3] Detailed derivations are provided in Appendix A.4.

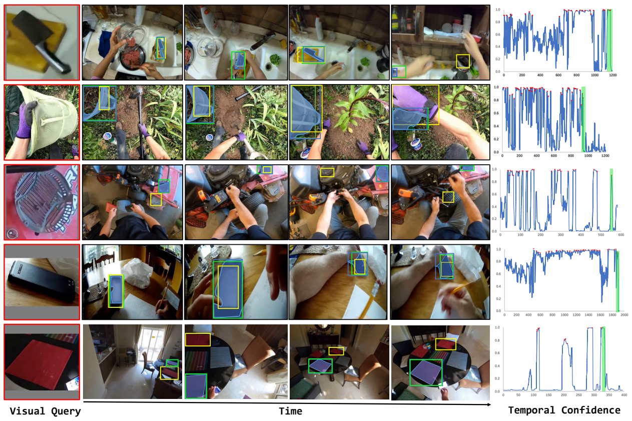


Figure 2: **Qualitative results in challenging scenarios.** Each row presents the visual query, four frames from the video, the predicted trajectories from both EAGLE and VQLoc, and the ground truth. Additionally, the temporal confidence curve predicted by EAGLE is shown, with the green shaded region indicating the ground-truth interval.

added to $\mathcal{O}_{GLM}$ as dynamic snapshots, which are updated following a FIFO policy. The model updater, $\mathcal{U}$, serves as the core module for constructing the DCF. It takes $\mathcal{O}_{GLM} = \{(\mathcal{F}(\cdot)_i, G(\cdot)_i)\}_{i=1}^{n}$ as input, where $\mathcal{F}(\cdot)_i$ represents either the query or the result of 2D branch features, and $G(\cdot)_i$ is the corresponding Gaussian label. The objective is to learn a set of convolutional kernel weights, c, to construct the target model, formulated as $c = \mathcal{U}(\mathcal{O}_{GLM})$.

We employ a least-squares regression loss function to supervise the training of the DCF, defined as follows:

$$\mathcal{L}(c) = \frac{1}{|\mathcal{O}_{GLM}|} \sum_{(\mathcal{F}(\cdot),G(\cdot))\in\mathcal{O}_{GLM}} \|\mathcal{H}(\mathcal{H}_{\mathcal{J}}, G)\|^2 + \|\lambda c\|^2. \tag{3}$$

Here, λ is the regularization coefficient. $\mathcal{H}(\mathcal{H}_{\mathcal{J}}, G)$ represents the spatial residual between the predicted score $\mathcal{H}_{\mathcal{J}} = \mathcal{F}(\cdot) * c$ and the corresponding Gaussian label G. Drawing inspiration from the efficacy of Hinge loss in addressing data imbalance, we formulate the residual as:

$$\mathcal{H} = sw_G \cdot (\mathcal{S}_i\mathcal{H}_{\mathcal{J}} + (1 - \mathcal{S}_i)\max(0, \mathcal{H}_{\mathcal{J}}) - G_i) \tag{4}$$

Here, sw_G denotes a spatial weighting function contingent on the Gaussian label G, which assigns higher weights to positions proximate to the target's center and lower weights to those in the background. $\mathcal{S}_i$ represents the corresponding distinct binary masks, originating from the initial query only when $i = 0$, and otherwise from 2D retrieval results. This formulation precisely fits the target score G_i when $\mathcal{S}_i = 1$; employs the hinge branch when $\mathcal{S}_i = 0$, focusing solely on whether $\mathcal{H}_{\mathcal{J}} > 0$; and automatically learns the boundary distance of the target object when $0 < \mathcal{S}_i < 1$. This approach enables the loss function to smoothly transition between least squares and hinge based on the pixel's distance to the target center.

Previous works optimize DCF through a limited number of iteration steps(Danelljan et al. 2019; Bhat et al. 2020): $c^{(i+1)} = c^{(i)} - \beta\nabla\mathcal{L}(c^{(i)})$. We adopt a more refined optimization strategy—steepest descent iteration—to compute an adaptive step size. First, we perform a quadratic approximation at the current estimate $c^{(i)}$:

$$\begin{aligned}\mathcal{L}(c) \approx \hat{\mathcal{L}}(c) = \frac{1}{2}(c - c^{(i)})^{\top} PDS^{(i)}(c - c^{(i)}) \\ + (c - c^{(i)})^{\top}\nabla\mathcal{L}(c^{(i)}) + \mathcal{L}(c^{(i)})\end{aligned} \tag{5}$$

Here, both c and $c^{(i)}$ are treated as vectors, and $PDS^{(i)}$ is a positive definite matrix. Subsequently, in the gradient direction (Eq.(5)), we solve for the step size β that minimizes the approximate loss by setting $\frac{\partial}{\partial\beta}\hat{\mathcal{L}}(c^{(i)} - \beta\nabla\mathcal{L}(c^{(i)})) = 0$, yielding:

$$\beta = \frac{\nabla\mathcal{L}(c^{(i)})^{\top}\nabla\mathcal{L}(c^{(i)})}{\nabla\mathcal{L}(c^{(i)})^{\top} PDS^{(i)}\nabla\mathcal{L}(c^{(i)})} \tag{6}$$

This formula computes the step size β in the update Eq.(5). Similar to (Danelljan et al. 2019; Bhat et al. 2019), we set $PDS^{(i)} = \frac{\partial^2\mathcal{L}}{\partial c^2}(c^{(i)})$, which is the Hessian matrix of the loss function (Eq.(3)), employing the second-order Taylor expansion from Eq.(6). For Eq.(3), the Gauss-Newton method based on first-order derivatives is more efficient, where $PDS^{(i)} = (J^{(i)})^{\top}J^{(i)}$, and $J^{(i)}$ is the Jacobian matrix of the residuals at $c^{(i)}$ [4]. Beyond the initial phase, during online inference, if the score map $\mathcal{H}_{\mathcal{J}}$ fails to consistently produce high responses (exceeding 60% of the historical frame length), we update the DCF using the original static snapshot; otherwise, we employ the most recent dynamic snapshot to update the model. This approach enables learning filter weights that balance static and dynamic scenarios, representing the responses of the query target against the background across varying times, scenes, and states via geometric response scores.

3.3 Dual Branches Integration

To mitigate the risk of individual branch degradation, we adopt a dual-branch framework wherein each branch provides complementary prior knowledge. Specifically, the tracking branch's score map, $\mathcal{H}_{\mathcal{J}}$, is encoded by a conv-bn-relu block, κ_θ, resulting in $\mathcal{F}_{\mathcal{J}} = \kappa_\theta(\mathcal{H}_{\mathcal{J}})$, which is dimensionality-matched to the mask encoding, $\mathcal{F}_{\mathcal{A}}$. These features are then fused via element-wise addition, yielding a combined feature representation, $\mathcal{F}_{\mathcal{N}} = \mathcal{F}_{\mathcal{A}} + \mathcal{F}_{\mathcal{J}}$. The decoder, $\mathbf{D}$, processes $\mathcal{F}_{\mathcal{N}}$ to produce a segmentation score map, $\mathbf{D}(\mathcal{F}_{\mathcal{N}})$. Finally, mask generation yields the segmentation result, $\mathcal{S}$, at the original resolution, and the bbox trajectory, $\mathcal{R}$, used for evaluation is derived from the minimum bounding rectangle enclosing the connected components of $\mathcal{S}$.

3.4 VGGT-Powered 3D Visual Localization

Given an input video sequence $\mathcal{V} = \{I_i\}_{i=1}^{N}$ consisting of N frames, VGGT jointly infers per-frame geometric cues in a single, end-to-end pass. For each image I_i, the model outputs the camera parameters $\mathcal{C}_i = \{\mathcal{T}_i, \mathcal{K}_i\}$, a dense depth

[4] Complete details are provided in Appendix A.2

Table 1: **Comparison results on Ego4D-VQ2D benchmark**. Ego4D provides the complete definitions of the evaluation metrics. † denotes the approach fine-tuned on Ego-Tracks, the same applies hereinafter.

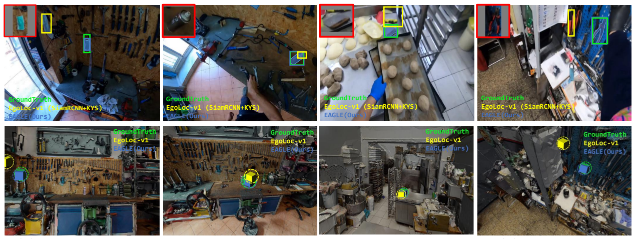
Method	VQ2D Test Server Leaderboard				VQ2D Validation Set			
	$\text{tAP}_{25}\uparrow$	$\text{stAP}_{25}\uparrow$	Rec.(%)↑	Succ.(%)↑	$\text{tAP}_{25}\uparrow$	$\text{stAP}_{25}\uparrow$	Rec.(%)↑	Succ.(%)↑
Ego4D baseline [CVPR'22]	0.20	0.13	32.20	39.80	0.22	0.15	32.92	43.24
NFM [Ego4D 2022 Winner]	0.24	0.17	35.29	43.07	0.26	0.19	37.88	47.90
CocoFormer [CVPR'23]	0.25	0.18	42.34	48.37	0.26	0.19	37.67	47.68
STARK-S50 [ICCV'21]	-	-	-	-	0.08	0.03	11.35	15.08
STARK-S50(†)	-	-	-	-	0.29	0.20	35.57	45.20
STARK-S101	-	-	-	-	0.10	0.04	12.41	18.70
STARK-S101(†)	-	-	-	-	0.30	0.21	41.11	48.03
VQLoc [NeurIPS'23]	0.32	0.24	45.10	55.88	0.31	0.22	47.05	55.89
RELOCATE [CVPR'25]	0.43	0.35	50.60	60.10	0.41	0.33	50.50	58.03
PRVQL [Ego4D 2025 Challenger]	0.37	0.28	45.70	59.43	0.35	0.27	47.87	57.93
EAGLE (Ours)	**0.63**	**0.61**	**53.51**	**62.70**	**0.60**	**0.58**	**52.09**	**61.29**

Table 2: **Comparison results on Ego4D-VQ3D benchmark**. * indicates official improved baseline.

Method	**VQ3D Test Server Leaderboard**					**VQ3D Validation Set**				
	Succ.(%)↑	**Succ.*(%)↑**	**L2↓**	**Angle↓**	**QwP(%)↑**	**Succ.(%)↑**	**Succ.*(%)↑**	**L2↓**	**Angle↓**	**QwP(%)↑**
Ego4D[CVPR'22]	7.95	48.61	4.64	1.31	0.16	-	-	-	-	-
Ego4D*[CVPR'22]	8.71	51.47	4.93	1.23	15.15	1.22	30.77	5.98	1.60	1.83
Eivul[Ego4D 2022 Challenger]	25.76	38.74	8.97	1.21	66.29	73.78	91.45	2.05	0.82	80.49
CocoFormer[CVPR'23]	9.09	50.60	4.23	1.23	16.29	-	-	-	-	-
EgoCOL[Ego4D 2023 Winner]	62.88	85.27	2.37	0.53	74.62	59.15	93.39	2.31	0.58	63.42
EgoLoc[ICCV'23]	87.12	96.14	1.86	0.92	90.53	80.49	98.14	1.45	0.61	82.32
EgoLoc-v1[CVPR'24]	88.64	96.15	1.86	1.24	92.05	81.13	98.10	1.45	0.55	84.73
EAGLE(Ours)	**89.02**	**96.14**	**1.84**	1.21	**92.42**	**84.77**	**98.54**	**1.18**	**0.42**	**85.68**

map $\mathcal{D}_i$, and a pixel-wise depth uncertainty map $\mathcal{UD}_i$. This pipeline reduces the processing time from the several minutes or even hours required by COLMAP to only a few seconds, while delivering geometry estimations of comparable accuracy. The resulting geometric priors form the backbone of the subsequent aggregation strategy. VGGT predicts 3D geometry in a self-consistent, yet arbitrary, coordinate system. To evaluate on the VQL-3D benchmark, we must align our predictions with the ground-truth matterport scan coordinate system for each sequence. This is achieved by solving for a 7-DoF similarity transformation, $\mathbf{T}_\eta \in \text{Sim}(3)$, which maps our predicted point cloud ($\mathbf{pc}^{\text{vggt}}$) to the ground-truth ($\mathbf{pc}^{\text{ms}}$). We formulate this as a least-squares optimization problem: $\mathbf{T}_\eta = \underset{\mathbf{T}\in\text{Sim}(3)}{\arg\min} \sum_j ||\mathbf{T} \cdot \mathbf{pc}_j^{\text{vggt}} - \mathbf{pc}_j^{\text{ms}}||^2$, where j indexes corresponding 3D point pairs. Once the optimal transformation $\mathbf{T}_\eta$ is found, we apply it to all predicted camera poses and 3D structures, thereby unifying them into the canonical benchmark coordinate system for evaluation.

We propose a novel multi-view aggregation mechanism that fuses the semantic confidence from the VQL-2D branch with the geometric uncertainty derived from the VGGT, enabling robust estimation of the spatial location of the object. The strategy relies on the key assumption that the 3D location of the target remains relatively stable within a short observation window, an assumption we deem reasonable for the majority of human-object interaction scenarios within Ego4D. Specifically, the VQL-2D network returns a visual track $\mathcal{R}$ in the form of bounding boxes during the retrieval stage. The center coordinates (u_i, v_i) of the corresponding segmentation mask $\mathcal{S}_i$ are used as the 2D localization result for the target in that frame. To associate location information across multiple views, we lift these 2D coordinates into 3D space using the aforementioned aligned geometric information. Based on the principle of inverse projection from the standard pinhole camera model, the corresponding 3D coordinates $[\mathcal{X}_i, \mathcal{Y}_i, \mathcal{Z}_i]^T$ are computed as:$[\mathcal{X}_i, \mathcal{Y}_i, \mathcal{Z}_i, 1]^T = (\mathbf{T}_\eta \mathcal{T}_i)\mathcal{D}_i(u_i, v_i)\mathcal{K}_i^{-1}[u_i, v_i, 1]^T$, where $\mathcal{T}_i$ and $\mathcal{K}_i$ correspond to the camera extrinsics and intrinsics, respectively.



(a) Query: *face mask* (b) Query: *spray can* (c) Query: *slicing knife* (d) Query: *pincers*

Figure 3: **Visualization of 2D responses and 3D localization**. We back-projected the 2D response predictions and query locations into 3D space. Groundtruth, EAGLE, and the VQL-3D winner, EgoLoc-V1, were compared. We do not actually know the size and rotation of the 3D bbox during the prediction process. However, for visualization purposes, we utilize the size and rotation from the ground truth annotations and treat the predicted 3D location as the center of the 3D bbox.

Given that single-view localization results are susceptible to various factors, we propose a multi-view aggregation process to obtain an accurate and robust 3D location of the target. The core idea is to perform a weighted average of the 3D coordinates contributed by each view, where the weights are jointly determined by the quality of the segmentation results from the 2D branch and the reliability of the 3D reconstruction. To this end, we define a fused weight $\mathcal{FW}_i = s^i_{conf} \cdot g^i_{conf}$, which is obtained by multiplying two orthogonal confidence components: the semantic confidence s^i_{conf} and the geometric confidence g^i_{conf}. This weight attains a higher value only when both the 2D segmentation quality and the 3D geometric reconstruction are reliable, thus ensuring the robustness of the aggregation process. s^i_{conf} is used to evaluate the quality of the 2D segmentation results from VQL-2D. Since the VQL-2D retrieval results are directly reflected by the segmentation results, the semantic confidence comprehensively measures the localization clarity by considering the pixel probabilities within the segmentation mask $\mathcal{S}$. Specifically, we define three sub-metrics: the average probability $\mathbf{P}_{av}$, the maximum probability $\mathbf{P}_{max}$, and the average probability of pixels above a specific threshold λ, denoted as $\mathbf{P}_\lambda$. These are defined as follows:

$$\begin{cases} \mathbf{P}_{av} = \frac{1}{|\mathcal{S}_i|} \sum_{(u,v)\in\mathcal{S}_i} pr_i^{(u,v)}, \\ \mathbf{P}_\lambda = \frac{1}{n} \sum_{(u,v)\in\mathcal{S}_i,\ pr_i^{(u,v)}>\lambda} pr_i^{(u,v)}, \\ \mathbf{P}_{max} = \max_{(u,v)\in\mathcal{S}_i} pr_i^{(u,v)}, \end{cases} \tag{7}$$

where $pr_i^{(u,v)}$ is the predicted probability that pixel (u, v) belongs to the target, $|\mathcal{S}_i|$ is the total number of pixels in the mask, and n is the number of pixels within the mask whose probability exceeds the threshold λ. The final semantic confidence is a linear combination of these three metrics: $s^i_{conf} = \varphi\mathbf{P}_{av} + \psi\mathbf{P}_\lambda + \mu\mathbf{P}_{max}$, where φ, ψ, and μ are hyperparameters, all set to 1/3 in our experiments. The geometric confidence $g_{conf}(i)$ is directly derived from the predicted uncertainty of VGGT. We extract the depth uncertainty value $\tau_i = \mathcal{U}_i(u_i, v_i)$ at the target's center point

(u_i, v_i) and convert it into a normalized confidence score $g^i_{conf} = \exp(-\zeta\tau_i)$, where ζ is a hyperparameter used to adjust the influence of the uncertainty. This confidence ensures that the contribution of a point to the final aggregation is effectively suppressed when VGGT lacks confidence in its depth prediction for that point. We apply the fused weights $\mathcal{FW}_i$ to multi-view aggregation to obtain a final, aggregated 3D world coordinate for the target, denoted as $[\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{Z}}]^T$:

$$[\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{Z}}]^T = \frac{\sum_{i=1}^{N} \mathcal{FW}_i \cdot [\mathcal{X}_i, \mathcal{Y}_i, \mathcal{Z}_i]^T}{\sum_{i=1}^{N} \mathcal{FW}_i}. \tag{8}$$

Finally, we transform this aggregated result back into the camera coordinate system of each specific frame, thus obtaining a 3D relative displacement vector δ_i with respect to the current observing camera: $\delta_i = (\mathbf{T}_\eta \mathbf{T}_i)^{-1}[\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{Z}}, 1]^T = (\Delta\mathcal{X}, \Delta\mathcal{Y}, \Delta\mathcal{Z})$. δ_i represents the final 3D localization result output by our framework for the i-th frame, and it implicitly incorporates the consistency constraints derived from multi-view information.

4 Experiments

4.1 Implementation

Video clips undergo preprocessing, which includes uniform scaling to a 448 × 448 resolution via cropping along the longer side and subsequent zero-padding. A Laplacian operator with a window size of 100 is then applied to filter frames, enhancing their clarity. Both training and inference phases involve partitioning the videos into fixed-length clips [5]. The backbone comprises a pre-trained ViT(Oquab et al. 2023), fine-tuned on EgoTracks and subsequently frozen. The segmentation branch is trained using the Ego4D, EgoTracks(Tang et al. 2024), and VISOR(Darkhalil et al. 2022) datasets. For the Ego4D and EgoTracks, which provide only bounding box annotations, we leverage SAM to generate segmentation masks, treating them as ground truth. Within the AMM, the meta-learner employs $N^{train}_{init} = 10$ and $N^{train}_{update} = 3$ for the initial phase and historical frames, respectively. The memory bank size ($\mathcal{O}_{MAX}$) for $\mathcal{O}_{AMM}$ and $\mathcal{O}_{GLM}$ is set to 50 [6]. The AdamW is employed for 25,000 iterations, with a peak learning rate of 0.0025 and a weight decay of 0.05, using a linear learning rate scheduler with 2,500 warm-up iterations. We utilize the 1B version of VGGT. We train and evaluate on GTX4090 GPUs.

Training. We construct a training sequence $\mathcal{V}_{tr}$ by randomly sampling N frames. The first frame initializes the memory banks $\mathcal{O}_{AMM}$ and $\mathcal{O}_{GLM}$. Subsequently, we only update the segmentation branch parameters σ while keeping the tracking parameters frozen. The model is optimized with a total loss $\mathcal{L}_{total} = \mathcal{L}_{seg} + \rho\mathcal{L}_{tck}$, where $\mathcal{L}_{seg}$ is the Lovász loss, defined as $\sum_{j=1}^{J-1} \mathcal{L}_\sigma\big(\mathbf{D}(\mathcal{A}_\sigma^{j-1}(\mathcal{F}_j) + \kappa_\theta(c(\mathcal{F}_j))), \mathcal{M}_j\big)$, and $\mathcal{L}_{tck}$ is a hinge loss, formulated as $\sum_{j=1}^{J-1} \frac{1}{N_{iter}} \sum_{i=0}^{N_{iter}} \mathcal{L}_c(\mathcal{F}_i * c, G_i)$. ρ is weighting factor [7].

Inference. The video is first processed in clips, and the predictions are concatenated. The dual memory banks are initialized with the visual query, updated continuously for the first 100 historical frames, and then every 25 frames thereafter. The initial query sample is augmented using flipping, translation, and blurring. We use the mask confidence score, s_{conf}, as the temporal score. After applying a 5-frame median filter, we set a threshold at $0.8 \times \max(s_{conf})$. The last interval exceeding this threshold is output as the final temporal localization for VQL-2D.

4.2 Comparison to the state-of-the-art

We evaluate EAGLE on Ego4D-VQ, the unique publicly available benchmark for VQL. As shown in Table.1, On the test set, our method surpasses the previous best, RELOCATE, by 46.5% in tAP$_{25}$, 74.3% in stAP$_{25}$, 5.8% in Rec, and 4.3% in Succ. On the validation set, we achieve performance gains of 46.3%, 75.8%, 3.1%, and 5.6% in the same metrics, respectively, while maintaining a comparable inference speed [8]. Qualitative comparisons with VQLoc, presented in Figure.2, highlight EAGLE's superior performance, robustness, and generalization capabilities in challenging egocentric scenarios. For the VQL-3D, EAGLE's performance is detailed in Table.2. Compared to EgoLoc-V1, EAGLE improves Succ and QwP by 0.4% on the test set, while reducing L2 and Angle errors by 1.1% and 2.4%. The improvements on the validation set are more substantial, with gains of 4.5% in Succ, 0.4% in Succ*, and 1.12% in QwP, alongside significant reductions of 18.6% in L2 and 23.6% in Angle errors. These advancements are attributed to VGGT's comprehensive estimation of camera pose and depth, which allows for the utilization of a greater number of camera poses in the computation. Qualitative results are visualized in Figure3.

4.3 Ablation Analysis

Impact of AMM. To investigate the impact of each component within AMM, we conduct an ablation study with five variants, as detailed in Table.3: (i) $STA \rightarrow SAM$: We replace SAM [9] with a pre-trained STA(Zhao et al. 2021) model during the pseudo-mask generation phase. (ii) w/o $\mathcal{P}_\theta$: We ablate the pseudo label modulator network, utilizing only binary mask information. (iii) w/o $\mathcal{W}_\theta$: We remove the target re-weighting network. (iv) $\mathcal{Q}_0! \rightarrow \phi$: Within $\mathcal{O}_{AMM}$, all queues except for the initial query's are set to a FIFO scheme, preventing the initial query from being replaced. (v) w/o $\mathcal{O}_{AMM}$: We completely remove the $\mathcal{O}_{AMM}$ module. The results indicate that the removal of any component adversely affects performance. Most notably, ablating the $\mathcal{O}_{AMM}$ leads to the most significant performance degradation, with drops of 10.9% in tAP$_{25}$, 13.6% in stAP$_{25}$, 24.3% in Rec, and 12.1% in Succ. This underscores the critical role of the memory bank in stabilizing target retrieval. The second most impactful change is the substitution of SAM with STA, which demonstrates that SAM generates higher-quality pseudo-labels containing more discriminative infor-

[5] Partitioning strategy in Appendix A.6

[6] Memory capacity ablations in Appendix A.6

[7] The hyperparameter configuration is detailed in Appendix A.6.

[8] Comparison of inference speed in Appendix A.6.

[9] Detailed settings for SAM is provided on Appendix A.6

Table 3: **Ablation study of AMM on the Ego4D-VQ2D validation set.** The final model configuration is highlighted in gray, a convention adopted hereinafter.

STA →SAM	w/o $\mathcal{P}_\theta$	w/o $\mathcal{W}_\theta$	$\mathcal{Q}_0$!→ϕ	w/o $\mathcal{O}_{AMM}$	Ego4D-VQ2D validation set			
					tAP$_{25}$↑	stAP$_{25}$↑	Rec.(%)↑	Succ.(%)↑
✓	-	-	-	-	0.57	0.46	45.19	57.28
-	✓	-	-	-	0.59	0.49	51.08	60.22
-	-	✓	-	-	0.60	0.48	51.55	61.01
-	-	-	✓	-	0.59	0.48	50.61	60.57
-	-	-	-	✓	0.55	0.44	42.22	54.62
-	-	-	-	-	**0.60**	**0.58**	**52.09**	**61.29**

Table 4: **Ablation study of GLM**

$\mathcal{Q}_0$ → ϕ	w/o $\mathcal{O}_{GLM}$	Ego4D-VQ2D validation Set			
		tAP$_{25}$↑	stAP$_{25}$↑	Rec.(%)↑	Succ.(%)↑
✓	-	0.57	0.45	46.35	57.73
-	✓	0.55	0.44	42.36	53.09
-	**-**	**0.60**	**0.58**	**52.09**	**61.29**

Table 5: **Ablation study of backbone and input resolution**

Backbone@Resolution	Ego4D-VQ2D validation Set			
	tAP$_{25}$ ↑	stAP$_{25}$ ↑	Rec.(%)↑	Succ.(%)↑
DINOv2-ViT-S/14@224	0.51	0.42	46.69	50.29
DINOv2-ViT-B/14@224	0.52	0.43	47.25	51.21
DINOv2-ViT-B/14@448	**0.60**	**0.58**	**52.09**	**61.29**
CLIP-ViT-B/16@448	0.54	0.45	49.96	52.45

Table 6: **Ablation study of multi-view aggregation function on the Ego4D-VQ3D validation set.**

Last-Resp. (Baseline)	s_{conf}			g_{conf}	Ego4D-VQ3D Validation Set				
	$\mathbf{P}_{av}$	$\mathbf{P}_\lambda$	$\mathbf{P}_{max}$		Succ.(%)↑	Succ.*(%)↑	L2↓	Angle↓	QwP%↑
✓	-	-	-	-	79.26	92.35	1.54	0.65	85.68
-	✓	-	-	-	79.95	92.76	1.52	0.63	85.68
-	-	✓	-	-	80.01	93.25	1.48	0.57	85.68
-	-	-	✓	-	81.65	94.53	1.34	0.51	85.68
-	✓	✓	-	-	81.92	96.75	1.28	0.49	85.68
-	✓	-	✓	-	82.22	96.49	1.25	0.45	85.68
-	-	✓	✓	-	82.29	96.88	1.26	0.45	85.68
-	-	-	-	✓	80.33	94.65	1.51	0.55	85.68
-	✓	✓	✓	✓	**84.77**	**98.54**	**1.18**	**0.42**	**85.68**

mation for retrieval, thereby enabling more precise segmentation.

Impact of GLM. We also designed two variants to investigate the impact of the GLM: (i) $\mathcal{Q}_0 \rightarrow \phi$: In $\mathcal{O}_{GLM}$, all queues are updated using a FIFO scheme, allowing the initial query to be replaced; and (ii) w/o $\mathcal{O}_{GLM}$: The $\mathcal{O}_{GLM}$ is completely removed. The experimental results show that removing $\mathcal{O}_{GLM}$ has the most pronounced impact on performance, leading to decreases of 10.9% in tAP$_{25}$, 13.6% in stAP$_{25}$, 23.8% in Rec, and 15.3% in Succ. This demonstrates that the historical visual cues provided by $\mathcal{O}_{GLM}$ is crucial for stable, long-term tracking within the discriminative branch. Interestingly, we find that the optimal update strategy for the GLM's memory is the inverse of that for the AMM; updating GLM's initial query along with the other memory sequences degrades performance. We attribute this to the distinct feature granularities handled by the two branches. The segmentation branch processes pixel-level information and is highly sensitive to contour and scale consistency; any accumulated error directly degrades the mask, so memory must be updated synchronously to reflect the latest appearance. In contrast, the GLM-guided discriminative tracking branch operates on region-level geometric descriptors and focuses on the target's identity ("what it is"), which gives it greater tolerance to minor appearance drift. Retaining the initial query as a fixed identity anchor effectively mitigates drift caused by prolonged occlusion or appearance changes, thereby maintaining tracking stability.

Impact of backbone size and input resolution. We conducted four variants:(i) DINOv2-ViT-S/14 with an input resolution of 448×448; (ii) DINOv2-ViT-B/14 with 224×224; (iii) DINOv2-ViT-B/14 with 448×448; and (iv) CLIP-ViT-B/16 [10] with 448×448. The results indicate that both backbone capacity and input resolution correlate positively with performance, whereas lowering the resolution leads to a substantial decline in localization accuracy. We therefore adopt the best-performing model, DINOv2-ViT-B/14@448.

Impact of multi-view aggregation function. Table.6 presents the ablation study on different variants of semantic and geometric confidence. The results indicate that each semantic confidence metric ($\mathbf{P}_{av}$, $\mathbf{P}_\lambda$, and $\mathbf{P}_{max}$), when used individually, outperforms the baseline model(only use last response). This superiority stems from their distinct approaches to handling mask pixel confidence: $\mathbf{P}_{av}$ mitigates the impact of missed detections by comprehensively evaluating the quality of region coverage; $\mathbf{P}_\lambda$ reduces background interference by filtering out low-confidence pixels via a threshold; and $\mathbf{P}_{max}$ focuses on high-confidence pixels to decrease sensitivity to noise and false positives, thereby significantly improving the L2 and Angle metrics. Notably, these three methods are complementary when combined. Furthermore, geometric confidence reflects the reliability of pixels in 3D space. Its integration with semantic confidence creates a synergistic effect, where their complementary strengths jointly enhance the precision of spatial localization.

5 Conclusions

In this paper, we addressed critical limitations in VQL by moving beyond conventional "detect-then-track" methods. We introduced a novel framework inspired by biological memory consolidation, which synergizes segmentation and tracking to build a robust, long-term episodic memory of the target. We establish the online memory bank that actively filters and stores high-confidence samples, enabling stable retrieval through significant appearance variations and environmental distractors. Furthermore, we unified 2D and 3D localization into an efficient pipeline. Our work achieves SOTA performance on the Ego4D-VQ benchmark, demonstrating its potential as a new baseline.

[10] https://github.com/openai/CLIP

References

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4413–4421.

Bevandić, J.; Chareyron, L. J.; Bachevalier, J.; Cacucci, F.; Genzel, L.; Newcombe, N. S.; Vargha-Khadem, F.; and Ólafsdóttir, H. F. 2024. Episodic memory development: Bridging animal and human research. *Neuron*, 112(7): 1060–1080.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6182–6191.

Bhat, G.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2020. Know your surroundings: Exploiting scene information for object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, 205–221. Springer.

Bolme, D. S.; Beveridge, J. R.; Draper, B. A.; and Lui, Y. M. 2010. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 2544–2550. IEEE.

Choi, J.; Kwon, J.; and Lee, K. M. 2019. Deep Meta Learning for Real-Time Target-Aware Visual Tracking. arXiv:1712.09153.

Dai, K.; Zhang, Y.; Wang, D.; Li, J.; Lu, H.; and Yang, X. 2020. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6298–6307.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4660–4669.

Darkhalil, A.; Shan, D.; Zhu, B.; Ma, J.; Kar, A.; Higgins, R.; Fidler, S.; Fouhey, D.; and Damen, D. 2022. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35: 13745–13758.

Dickerson, B. C.; and Eichenbaum, H. 2010. The episodic memory system: neurocircuitry and disorders. *Neuropsychopharmacology*, 35(1): 86–104.

Fan, B.; Feng, Y.; Tian, Y.; Lin, Y.; Huang, Y.; and Fan, H. 2025. PRVQL: Progressive Knowledge-guided Refinement for Robust Egocentric Visual Query Localization. *arXiv preprint arXiv:2502.07707*.

Forigua, C.; Escobar, M.; Pont-Tuset, J.; Maninis, K.-K.; and Arbeláez, P. 2023. EgoCOL: Egocentric Camera pose estimation for Open-world 3D object Localization@ Ego4D challenge 2023. *arXiv preprint arXiv:2306.16606*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.

Hao, S.; Chai, W.; Zhao, Z.; Sun, M.; Hu, W.; Zhou, J.; Zhao, Y.; Li, Q.; Wang, Y.; Li, X.; et al. ???? Ego3DT: Tracking Every 3D Object in Ego-centric Videos. In *ACM Multimedia 2024*.

Jiang, H.; Ramakrishnan, S. K.; and Grauman, K. 2024. Single-stage visual query localization in egocentric videos. *Advances in Neural Information Processing Systems*, 36.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Khosla, S.; Schwing, A.; Hoiem, D.; et al. 2025. Relocate: A simple training-free baseline for visual query localization using region-based representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3697–3706.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lukezic, A.; Matas, J.; and Kristan, M. 2020. D3s-a discriminative single shot segmentation tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7133–7142.

Mai, J.; Hamdi, A.; Giancola, S.; Zhao, C.; and Ghanem, B. 2023. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 45–57.

Mai, J.; Hamdi, A.; Giancola, S.; Zhao, C.; and Ghanem, B. 2024. Hybrid Structure-from-Motion and Camera Relocalization for Enhanced Egocentric Localization. *arXiv preprint arXiv:2407.08023*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Park, E.; and Berg, A. C. 2018. Meta-Tracker: Fast and Robust Online Adaptation for Visual Object Trackers. arXiv:1801.03049.

Plizzari, C.; Goletto, G.; Furnari, A.; Bansal, S.; Ragusa, F.; Farinella, G. M.; Damen, D.; and Tommasi, T. 2024. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 1–57.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Tang, H.; Liang, K. J.; Grauman, K.; Feiszli, M.; and Wang, W. 2024. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 36.

Tschernezki, V.; Larlus, D.; and Vedaldi, A. 2021. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *2021 International Conference on 3D Vision (3DV)*, 910–919. IEEE.

Ullman, S. 1979. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153): 405–426.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Xu, M.; Fu, C.-Y.; Li, Y.; Ghanem, B.; Perez-Rua, J.-M.; and Xiang, T. 2022. Negative Frames Matter in Egocentric Visual Query 2D Localization. *arXiv preprint arXiv:2208.01949*.

Xu, M.; Li, Y.; Fu, C.-Y.; Ghanem, B.; Xiang, T.; and Pérez-Rúa, J.-M. 2023. Where is my wallet? modeling object proposal sets for egocentric visual query localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2593–2603.

Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; and Lin, L. 2019. Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhao, B.; Bhat, G.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2021. Generating masks from boxes by mining spatio-temporal consistencies in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13556–13566.

Zhao, Y.; Ma, H.; Kong, S.; and Fowlkes, C. 2024. Instance Tracking in 3D Scenes from Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21933–21944.