Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Egocentric Object-Interaction Anticipation with Retentive and Predictive Learning

Guo Chen[1] , **Yifei Huang**[2] , **Yin-Dong Zheng**[1] , **Yicheng Liu**[1] , **Jiahao Wang**[3] , **Tong Lu**[1]
[1]Nanjing University
[2]The University of Tokyo
[3]Kuaishou Technology
chenguo1177@gmail.com, hyf@iis.u-tokyo.ac.jp, ydzheng0331@gmail.com,
lyccnb@gmail.com, wangjiahao08@kuaishou.com, lutong@nju.edu.cn

Abstract

Egocentric object-interaction anticipation is critical for applications like augmented reality and robotics, but existing methods struggle with misaligned egocentric encoding, insufficient supervision, and underutilized historical context. These limitations stem from a lack of focus on retention, *i.e.*, retaining long-term object-centric interactions, and prediction, *i.e.*, future-centric encoding and future uncertainty modeling. We introduce EgoAnticipator, a novel Retentive and Predictive Learning framework that addresses these challenges. Our approach combines retentive pre-training for domain-specific encoding, predictive pre-training for future uncertainty modeling, and mirror distillation to transfer future-informed knowledge. Additionally, we propose long-term memory prompting to integrate historical interaction cues. We evaluate the effectiveness of our framework using the Ego4D short-term object interaction anticipation benchmark, covering both STAv1 and STAv2. Extensive experiments demonstrate that our framework outperforms existing methods, while ablation studies highlight the effectiveness of each design inside our retentive and predictive learning framework.

1 Introduction

Recent advances in wearable camera technologies facilitate the collection of large-scale egocentric video datasets [Damen *et al.*, 2018; Huang *et al.*, 2024; Grauman *et al.*, 2022], thereby significantly boosting the research in egocentric video analysis. While the main focus of research is on recognition tasks such as action recognition [Wang *et al.*, 2021; Patrick *et al.*, 2021; Núñez-Marcos *et al.*, 2022] and hand grasp analysis [Cai *et al.*, 2016], the anticipation of future human-object interaction is receiving increasing attention [Grauman *et al.*, 2022; Thakur *et al.*, 2024]. This increased focus is largely driven by its potential applications, particularly in VR/AR [Wang *et al.*, 2023; Huang *et al.*, 2018], where, for example, AR assistants can proactively suggest upcoming actions, predict upcoming movements in telesurgery to prepare robotic systems.

The task of anticipating egocentric object interaction aims to predict the future state of human-object interaction without directly observing future videos. This includes forecasting the object's position and category, the form of interaction, and the timing of the interaction. A straightforward solution is to extend the current spatio-temporal localization problem [Köpüklü *et al.*, 2019; Murray *et al.*, 2012] to fit this anticipation task. These methods [Grauman *et al.*, 2022; Ragusa *et al.*, 2023] focus on analyzing the current input to enhance future interaction anticipations. However, these methods often underperform due to three key limitations.

Firstly, there is a *misalignment in egocentric encoding*. Existing methods rely on pre-training on general video datasets with mixed viewpoints, such as Kinetics [Kay *et al.*, 2017]. This results in encoders that lack domain-specific alignment for egocentric perspectives. In addition, these encoders are typically trained with descriptive labels that emphasize ”what is happening”, which restricts their ability to encode the **predictive** context in the video representations.

Secondly, *supervision signals are insufficient*. In real-world scenarios, future events are probabilistic rather than deterministic. However, datasets often employ one-hot hard labels, assuming a single outcome with absolute certainty. This undermines the effectiveness of supervision and introduces bias, limiting the models' **predictive** ability in capturing the inherent uncertainty and variability of future events.

Thirdly, there is *insufficient utilization of historic information*. Existing methods [Ragusa *et al.*, 2023; Mur-Labadia *et al.*, 2024] typically focus solely on short video clips without leveraging the **retention** of long-term historical data. This oversight can lead to the omission of crucial context that could significantly enhance anticipation accuracy.

To address these limitations, in this work, we propose a **Retentive and Predictive Learning** framework, named EgoAnticipator, specifically designed for egocentric object-interaction anticipation. Our framework aims to mitigate the above three issues.. To formulate pertinent memory encoding that is effective in the anticipation context, we first conduct a *retentive and predictive pre-training*. This step encourages the model to retain important cues of human-object interaction and transfer the cues into a format conducive to anticipation. Following this, we implement a *mirror distillation* strategy, which utilizes a teacher model with superior future insight, to assist in both the formulation of memory and the ex-

Preprint

ecution of the anticipation task. This teacher model is trained on future frames that are mirror-flipped relative to the current timestamp, aiding in memory formulation. Moreover, it introduces a level of predictive uncertainty, which is critical for modeling anticipation tasks. Thirdly, to effectively retrieve pertinent object interaction information from the memory, we craft a *long-term memory prompting* method. We formulate a prompting feature from long-term history using the object attributes generated by the student or teacher model. Since object attributes are critical in the object interaction anticipation task, long-term memory prompting can effectively enhance the model's anticipatory capability.

We evaluate the performance of our method on the Ego4D-STA benchmark [Grauman *et al.*, 2022]. EgoAnticipator effectively addresses the challenges of domain-specific encoding, probabilistic supervision, and long-term information utilization, resulting in significantly enhanced prediction accuracy. Through comprehensive evaluations, our framework consistently outperforms existing state-of-the-art methods, demonstrating its effectiveness and robustness in anticipating human-object interactions from egocentric perspectives.

2 Related Work

2.1 Action and Object-Interaction Anticipation

Prior works [Furnari and Farinella, 2020; Girdhar and Grauman, 2021; Xu *et al.*, 2021; Sener *et al.*, 2020] in action anticipation mainly focus on classifying future actions within a preset time frame. Approaches [Zhao and Krähenbühl, 2022; Furnari and Farinella, 2020; Xu *et al.*, 2021; Sener *et al.*, 2020] delve into temporal modeling strategies for forecasting forthcoming events at predetermined timestamps. Diverging from the typical temporal action anticipation tasks with fixed intervals, Ego4D introduces a novel short-term object-interaction anticipation task, requiring predictions about interactions with the next active objects [Pirsiavash and Ramanan, 2012; Furnari *et al.*, 2017] and time to contact them. This task has recently been further explored in studies such as [Ragusa *et al.*, 2023; Thakur *et al.*, 2024; Chen *et al.*, 2022; Mur-Labadia *et al.*, 2024]. In this work, we take a step forward by broadening the scope of information utilized in object interaction anticipation tasks. By incorporating our novel retentive and predictive learning framework, we can effectively establish historical memory and retrieve relevant object interaction from this memory for more accurate and nuanced anticipation about future interactions.

2.2 Knowledge Distillation (KD)

[Hinton *et al.*, 2015] has been established as an effective approach to transfer knowledge from a complex, high-capacity teacher model to a smaller, simpler student model. Its applications span diverse tasks, including image classification [Huang *et al.*, 2022], and dense prediction [Liu *et al.*, 2019]. In the typical implementation of KD [Chen *et al.*, 2017; Hinton *et al.*, 2015], the student model is trained to mimic the soft targets generated by the teacher model. Other methods [Chen *et al.*, 2017; Zhang and Ma, 2020] have focused on minimizing discrepancies between intermediate layer representations of the student and teacher models. KD's versatility extends into the various video domains [Li *et al.*, 2021; Zhao *et al.*, 2020; Mullapudi *et al.*, 2019] or action anticipation tasks [Tran *et al.*, 2021; Fernando and Herath, 2021; Furnari and Farinella, 2023] as well. Our key insight about STA is that the most valuable clues reside in future events. Our proposed mirror distillation utilizes the object frame as a demarcation point to build future teachers and current students. This approach significantly enhances the student's capability of learning anticipation from memory by providing better uncertainty estimates.

2.3 Long-term Memory Learning

The efficacy of context learning in language memory is well-established, as demonstrated by the success of large language models [Brown *et al.*, 2020; Min *et al.*, 2022; Dong *et al.*, 2023]. Similarly, video is also a form of sequential data that can benefit from extended contextual memory, as shown in recent studies [Xu *et al.*, 2021; Wu *et al.*, 2022; Gao and Wang, 2023]. For example, MeMViT [Wu *et al.*, 2022] enhances MViT [Fan *et al.*, 2021] by incorporating long-term multi-scale memory prompts. In our work, we leverage the future teacher in our framework to retroactively update interaction and contact timings of past objects, retaining a refined object-interaction memory that provides more reliable contextual cues for future anticipation.

3 Preliminaries

To illustrate our method clearly, we first define the STA problem and describe the baseline model.

3.1 Problem Definition

We define the STA task as proposed in [Grauman *et al.*, 2022]. Given an input video V and a timestamp t, the input to the model is the video from its beginning up to time t (denoted as $V_{1:t}$). The model is required to anticipate when and where, and what kind of object interaction will happen. The predictions at time t can be represented as $\Psi = \{\varphi_i = (b, n, v, \delta, p)\}_{i=1}^{N_p}$, where b denotes the spatial location of object box, n is the object's category, while v, δ, and p denote the anticipated action, the time to contact (TTC) and the confidence score used to rank predictions for evaluation, respectively.

3.2 Baseline

We use the two-stage baseline from Ego4D [Grauman *et al.*, 2022], which we build upon in the next section. The stages are object detection and anticipation.

Stage1: Object Detection. In the first stage, an object detector such as Faster R-CNN [Girshick, 2015] generates multiple potential next-active objects per frame, resulting in predictions represented as $\tilde{\Psi} = \{\varphi_i = (b, \tilde{n}, \tilde{p})\}_{i=1}^{N_p}$, where b, $\tilde{n}$ and $\tilde{p}$ are the location, noun category and score of object predicted by object detector. These serve as inputs for the second stage.

Stage2: Short-term Anticipation. In the second stage, a video encoder processes low-resolution clips $V_{t-l:t}$ as short-term memory, covering l history frames and ending at timestamp t. The encoder extracts dynamic features, capturing

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

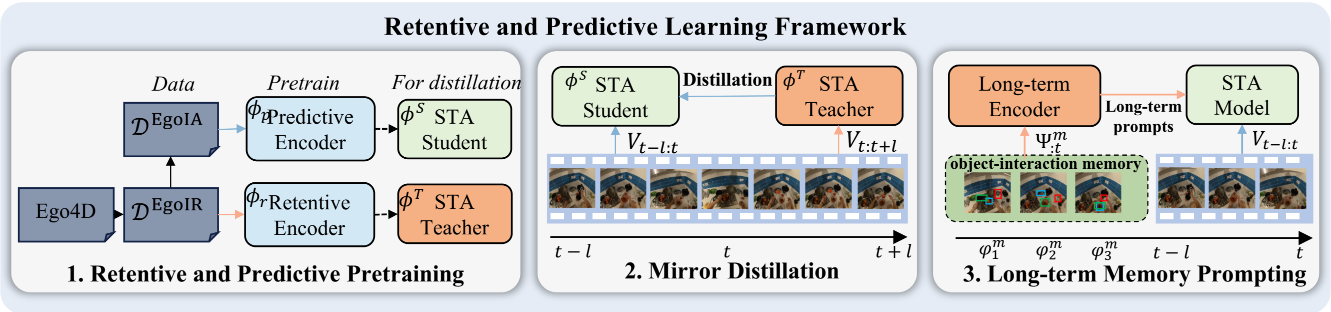

Figure 1: **Retentive and Predictive Learning Framework.** Our method curates relevant Ego4D data into $\mathcal{D}^{\text{EgoIR}}$ and $\mathcal{D}^{\text{EgoIP}}$ to train the Retentive Encoder ϕ_r and Predictive Encoder ϕ_a. These encoders are used by the teacher ϕ^T and student ϕ^S in *Retentive and Predictive Pre-training*. Subsequently, the student ϕ^S and teacher ϕ^T receive historical frames $V_{t-l:t}$ and future frames $V_{t:t+l}$, respectively, for Mirror Distillation. Meanwhile, ϕ^S infers long-term samples $V_{1:t}$, generating long-term memory prompts to enhance information. Finally, the STA Model, supported by multiple inputs, predicts object interactions including nouns n, verbs v, and time-to-contact δ (ttc).

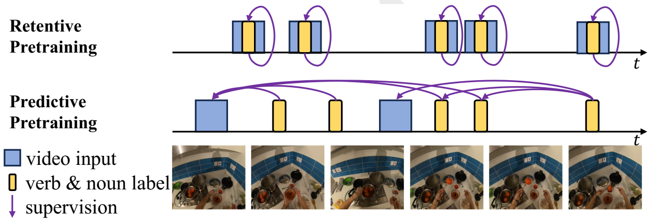

Figure 2: Illustration of **Retentive and Predictive Pretraining**.

the evolving actions and interactions. The video features $\mathbf{f}_{3d}$ and predictions $\tilde{\Psi}$ are fed into an anticipation network, which pools the spatial-temporal features to a 2D feature map $\mathbf{f}_{2d}$, then uses RoIAlign to sample region features $\mathbf{f}^b_{2d}$ for each bounding box b. These features $\mathbf{f}_b$ aid in forecasting the verb and time to contact. Finally, $\mathbf{f}_b$ is passed to classifiers to predict the object category n, verb v, and time to contact δ. The loss function combines cross-entropy loss $\mathcal{L}_{CE}$ on n and v with smooth L1 loss $\mathcal{L}_{L1}$ on δ, with $\lambda_1 = 10$:

$$\mathcal{L}_{sta} = \mathcal{L}^n_{CE} + \mathcal{L}^v_{CE} + \lambda_1 \mathcal{L}^\delta_{L1}. \tag{1}$$

4 Retentive and Predictive Learning

We introduce EgoAnticipator, a novel retentive and predictive learning framework for short-term object interaction anticipation in egocentric videos. As shown in Figure 1, it enhances the two-stage baseline from Section 3 with three key components: retentive and predictive pre-training, mirror distillation, and long-term memory prompting.

4.1 Retentive and Predictive Pre-training

A direct option to encode short-term memory with short video input $V_{t-l:t}$ is to use features extracted from a pre-trained video model. However, such models often lack domain-specific alignment for egocentric perspectives and are optimized for descriptive tasks rather than predictive ones. Thus, we craft **Retentive and Predictive Pre-training (RPP)**, a simple yet effective process to optimize the model's ability to retain and anticipate egocentric object interactions, as shown in Figure 2.

Retentive Pretraining. The first part of RPP focuses on training the video encoder ϕ_r to retain short-term egocentric memory, emphasizing object-centric interactions. We utilize the narration annotations from Ego4D as our training data. Following EgoVLP [Lin *et al.*, 2022], we extract verbs and nouns from each timestamp narration and extend their context boundaries. To ensure the relevance of our data to egocentric interactions, we selectively filter out entries marked with ”#o” and ”#O”, as they relate to exocentric information. The curated dataset, $\mathcal{D}^{\text{EgoIR}}$, encompasses 3.2M video clips featuring 115 verbs, 554 nouns, and forms the core dataset for training ϕ_r. Training ϕ_r on $\mathcal{D}^{\text{EgoIR}}$ is approached as a multi-task, multi-label classification task.

Predictive Pretraining. The second stage enhances the encoder's predictive capabilities by training ϕ_p on $\mathcal{D}^{\text{EgoIP}}$, a dataset derived from $\mathcal{D}^{\text{EgoIR}}$, using a sliding window of T_w seconds on each video. Clips are selected if a timestamped annotation occurs within the next T_f seconds, resulting in 6.5M clips annotated with future events. In $\mathcal{D}^{\text{EgoIP}}$, for each selected video window, future timestamp annotations are divided into N_f grids, each covering a fraction of $\frac{T_f}{N_f}$ seconds into the future. We employ N_f future queries and a transformer decoder block to decode future action sequences from each video window. The classification heads used in this stage are identical to those in retentive pre-training, ensuring consistency.

4.2 Mirror Distillation

To enhance the final anticipation model, we propose **Mirror Distillation**, which leverages a teacher model ϕ^T to provide (1) a reliable estimate of uncertainty for training ϕ^S, and (2) a true representation of short-term *future* features for enhancing ϕ^S via knowledge distillation. An overview of this step is illustrated in Figure 3.

Future Teacher. One primary limitation of the baseline model (ϕ^S without our enhancements) is its reliance solely on short-term memory $V_{t-l:t}$, limiting its ability to anticipate future interactions. To address this, we introduce a future teacher model ϕ^T with direct access to future data for guidance. The future teacher ϕ^T utilizes a mirror-flipping technique applied to video frame sampling around timestamp

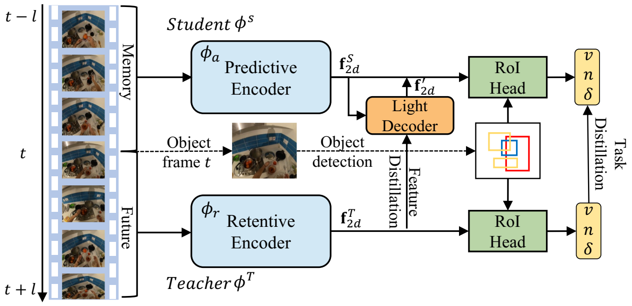

Figure 3: **Mirror distillation** goes beyond direct imitation of the teacher's output, by leveraging a residual connection to endow the student network ϕ^S with the capacity to assimilate the foresight of the teacher ϕ^T encoded in its representations.

t, allowing it to access short-term future information $V_{t:t+l}$, whereas the student model ϕ^S only accesses $V_{t-l:t}$. As illustrated in Figure 3, the teacher employs a retentive encoder ϕ_r, while the student uses a predictive encoder ϕ_p. We guide the learning of ϕ^S in two ways: (1) task distillation to provide reliable uncertainty estimation and (2) residual feature distillation to align phi^S's features more closely with future features.

Task Distillation. Task distillation aligns ϕ^T and ϕ^S by leveraging their shared context, *i.e.*, the center frame at t. Although they process different video segments, they both generate features $\mathbf{f}^T_b$ and $\mathbf{f}^S_b$ for the same objects in the center frame. We minimize the KL divergence between their predicted verb logits $\mathbf{y}^T_v$ and $\mathbf{y}^S_v$ using the following loss:

$$\mathcal{L}^v_{TD} = \frac{1}{N_p}\sum_t \tau^2 \mathcal{L}_{KL}(\sigma(\mathbf{y}^S_v/\tau), \sigma(\mathbf{y}^T_v/\tau)). \qquad (2)$$

where N_p is the number of objects, τ controls the softening, and σ is the Softmax function. Similarly, we define $\mathcal{L}^n_{TD}$ for nouns. We also distill knowledge related to time-to-contact (δ) prediction. Given δ^T and δ^S predicted by the teacher ϕ^T and the student ϕ^S, we minimize the smooth L1 loss $\mathcal{L}^\delta_{TD}$ between them by employ a separate MLP in ϕ^S for the distilled δ prediction. The teacher ϕ^T often generates higher-quality confidence scores p for each object, so we also distill knowledge from this score. We add an MLP to enable ϕ^S to learn the confidence score as fused by ϕ^T. The overall distillation loss for the 4 tasks can be represented as:

$$\mathcal{L}_{TD} = \mathcal{L}^v_{TD} + \mathcal{L}^n_{TD} + \lambda_2\mathcal{L}^\delta_{TD} + \mathcal{L}^p_{TD}, \qquad (3)$$

where λ_2 keeps consistent with λ_1.

Feature Residual Distillation. The task distillation transfers knowledge of individual objects from future to past frames, enhancing the student model's instance-level anticipation. To improve anticipation at a holistic level, we introduce a feature-level distillation.

Since ϕ^S and ϕ^T differ in their input data, directly enforcing the student's representation to be similar to the teacher's can result in the student model overly focusing on the future. This can cause the student model to overlook critical cues present in the current frames. To circumvent this issue, *Feature Residual Distillation* employs the residual connection for fusing predicted future features with current ones. Given the 2D feature map $\mathbf{f}^S_{2d}$ and $\mathbf{f}^T_{2d}$ at timestamp t, generated by ϕ^S and ϕ^T respectively, we utilize a lightweight feature decoder ψ within ϕ^S to predict future feature $\mathbf{f}'_{2d} = \psi(\mathbf{f}^S_{2d})$. We then minimize the smooth L1 loss between $\mathbf{f}'_{2d}$ and $\mathbf{f}^T_{2d}$, denoted as

$$\mathcal{L}_{FD} = \mathcal{L}_{L1}(\mathbf{f}'_{2d}, \mathbf{f}^T_{2d}). \qquad (4)$$

Subsequently, we update $\mathbf{f}^S_{2d}$ by fusing both features: $\mathbf{f}^S_{2d} = \mathbf{f}'_{2d} + \mathbf{f}^S_{2d}$. The newly updated features are then utilized in RoIAlign for the subsequent network.

4.3 Long-term Memory Prompting

While mirror distillation enhances near-term context encoding, it neglects long-term historical information. To this end, we propose **Long-term Memory Prompting** (depicted in Figure 4), to retrieve a broader range of object interaction information extended further back in time.

Long-term Object Interaction Representation. In the observed video $V_{1:t}$, we extract key object attributes for interaction anticipation: spatial location b, noun n, verb v, and contact time δ. The detected object-interaction memories are represented as $\Psi^m_{1:t} = \{\varphi^m_i = (b, n, v, p, d)\}^{N_p}_{i=1}$, where b is obtained by an object detector, and n, v, p are inferred by the memory encoder ϕ^M. Here, d denotes the temporal distance from the latest timestamp t. Each memory $\varphi^m_i = (b, n, v, \delta, d)$ is described as "At time $t-d$: after δ seconds, action v occurs on n at position b." We use separate MLPs to embed b, δ, and d, and learnable embeddings for n and v. These embeddings are concatenated for each object φ^m_i to form the representation $\mathbf{f}^m_i$, which is then input to the encoder of ϕ^S. There are two options for ϕ^M to create the long-term memory representation under different perspectives: 1) $\phi^M = \phi^S$: Use ϕ^S to represent the information before the historical target frame; 2) $\phi^M = \phi^T$: Use ϕ^T to represent the information after the historical target frame, as shown in Figure 5. The study of both options is further explained later in the next sections.

Long-term Memory Prompting. We incorporate a specialized lightweight network, denoted as ϕ_m and illustrated in Figure 4, to integrate the history of object interactions. The lightweight network consists of a single attention layer followed by an MLP, which encodes historical tuples into prompts for fusion with current-frame features. Its primary function is to encode the historical object interaction embeddings $\mathbf{f}^m_i$ into a long-term memory prompt, denoted by $\mathbf{f}^m_{\text{prompt}} = \phi_m(\mathbf{f}^m_i)$. This prompt assists in the final anticipation by fusing $\mathbf{f}^m_{\text{prompt}}$ with each object RoI feature $\mathbf{f}_b$ on the current frame t (see Sec. 3) through channel-wise concatenation followed by a fully-connected layer. The resulting fused object feature is then used to output the final anticipation result. Notably, during training with $\phi^M = \phi^T$, since ϕ^T requires the use of the l-frame context after f_{-1} frame, we only utilize objects that satisfy $d > l$ as accessible memory to prevent information leakage.

4.4 Multi-stage Training

The training schedule for EgoAnticipator consists of four steps: 1) Pre-train the video encoder ϕ on $\mathcal{D}^{\text{EgoIR}}$ to obtain

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

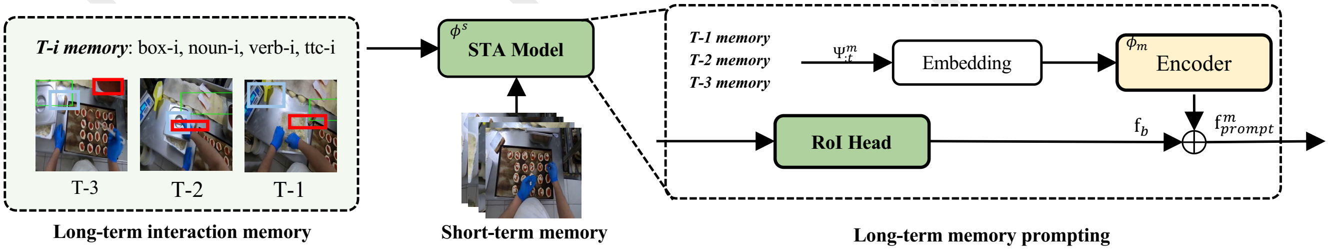

Figure 4: **Long-term memory prompts** emerge from the teacher's retroactive reasoning over long-term memory. For example, "**T-1 memory**: box-1, noun-1, verb-1, ttc-1" denotes the history "did *verb1* action on *noun1* at location *box1* after *ttc1* seconds". With history object interaction information embedded, the long-term prompt $\mathbf{f}^m_{prompt}$ is fused with the student network's RoI features to promote anticipation.

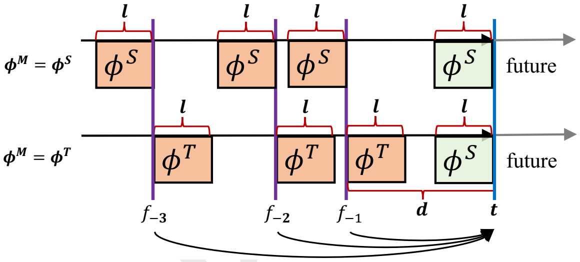

Figure 5: The illustration of training and inference of long-term memory prompting process with two options for ϕ^M.

Model (mAP)	b+n	b+v	b+n+t	b+n+v	b+n+v+t
One-stage training					
StillFast	16.20	-	4.94	7.47	2.48
Transfusion	20.19	-	6.17	7.55	2.60
STAformer	21.71	-	7.24	10.75	3.53
StillFast + EgoVLP	16.48	6.41	5.32	7.81	3.10
StillFast + $\mathcal{D}^{\text{EgoIR}}$	22.20	9.34	7.92	11.13	4.51
StillFast + $\mathcal{D}^{\text{EgoIP}}$	**23.01**	**9.68**	**8.53**	**11.66**	**4.84**
Two-stage training					
Slowfast	17.55	-	5.37	5.19	2.07
InternVideo	17.55	-	5.83	6.30	2.43
EgoAnticipator†	**18.34**	**10.67**	**7.72**	**8.03**	**4.29**
EgoAnticipator	**19.12**	**11.06**	**8.31**	**8.09**	**4.43**

Table 1: Results of our model and other baseline methods on STAv1. † denotes $\phi^M = \phi^S$.

ϕ_r. 2) Pre-train ϕ_r on $\mathcal{D}^{\text{EgoIP}}$ to acquire ϕ_p. 3) Use ϕ_p to train the teacher model ϕ^T, which observes an l-frame future, on $\mathcal{D}^{\text{STA}}$. 4) Train the student model ϕ^S with long-term memory prompts and perform mirror distillation with the teacher on $\mathcal{D}^{\text{STA}}$. The final loss for the STA student model combines three components:

$$\mathcal{L}_{stu} = \mathcal{L}_{sta} + \mathcal{L}_{TD} + \mathcal{L}_{FD}. \tag{5}$$

Additional details on the model architecture and the training of ϕ_r, ϕ_p, and ϕ^T are provided in the supplemental material.

4.5 Online Inference

The inference phase utilizes the trained student and teacher models. As illustrated in Figure 5, the orange and green regions represent the context ranges l fed into the model. We employ the student model ϕ^S to encode the short-term visual memory (green region) online and to aggregate the long-term interaction history queue ($Q = \{\dots, f_{-3}, f_{-2}, f_{-1}\}$) extracted by ϕ^M. As the visual input stream accumulates, the memory model ϕ^M continuously forms new interaction memories from the past (orange region). During inference with $\phi^M = \phi^S$, we directly append the already encoded short-term memory to the queue without recalculating it. However, when $\phi^M = \phi^T$, we must recalculate the new historical memory representation before inserting it into the queue. Additionally, we must ensure that the latest interaction memory f_{-1} (which may leak future information) is not involved in anticipating the current timestamp. To achieve this, we ensure that the time distance d between f_{-1} and t is greater than l by only appending f_{-1} to Q if $d > l$. This issue does not arise when $\phi^M = \phi^S$.

Model (mAP)	b+n	b+v	b+n+t	b+n+v	b+n+v+t
One-stage training					
StillFast	20.26	-	7.16	10.37	3.96
STAformer	24.85	-	7.41	13.45	4.90
StillFast + $\mathcal{D}^{\text{EgoIR}}$	25.75	**12.35**	9.30	14.41	5.56
StillFast + $\mathcal{D}^{\text{EgoIP}}$	**26.00**	11.92	**9.62**	**14.44**	**5.74**
Two-stage training					
Slowfast	21.00	-	7.04	7.45	2.98
EgoAnticipator†	**22.54**	**10.89**	**9.67**	**10.28**	**5.41**
EgoAnticipator	**23.52**	**11.18**	**10.29**	**11.04**	**5.60**

Table 2: Results of our model and other baseline methods on STAv2. † denotes $\phi^M = \phi^S$.

5 Experiments

5.1 Datasets and Evaluation Metric

We evaluate our framework on the **STAv1** and **STAv2** benchmarks from the Ego4D dataset. These benchmarks comprise 120 hours of annotated clips, including 27,801/98,276 training, 17,217/47,395 validation, and 19,780/19,780 test samples, spanning 87/128 noun and 74/81 verb classes. Evaluation employs Top-K AP/mAP, focusing on Top-5 metrics: Noun (b+n), Noun + Verb (b+n+v), Noun + TTC (b+n+δ), and Noun + Verb + TTC ($b + n + v + \delta$). This measures the model's ability to predict next-active object interactions at varying granularities.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Data	b+n	b+v	b+n+t	b+n+v	b+n+v+t
K400	16.94	7.61	6.37	4.81	2.34
$\mathcal{D}^{\text{EgoIR}}$	**17.25**	8.14	6.68	5.77	2.88
$\mathcal{D}^{\text{EgoIR}}$ w/ Exo	16.57	7.91	6.37	5.62	2.65
$\mathcal{D}^{\text{EgoIP}}$	17.07	**8.39**	**6.95**	**6.23**	**2.99**

Table 3: Impact of data of retentive and predictive pre-training.

Cls	TTC	Feat	Score	b+n	b+v	b+n+t	b+n+v	b+n+v+t
✓				18.50	9.36	7.45	6.80	3.20
✓	✓			18.27	9.01	7.75	6.81	3.34
✓	✓	✓		18.26	9.22	7.66	6.86	3.35
✓	✓	✓	✓	**18.67**	**9.28**	**7.87**	**6.92**	**3.45**

Table 4: Impact of distillation supervision.

Short	Long-i	Long-v	b+n	b+v	b+n+t	b+n+v	b+n+v+t
✓			17.37	8.39	7.01	5.83	2.99
✓	✓		**18.80**	**10.23**	**7.97**	**7.80**	**3.91**
✓		✓	18.18	9.69	7.38	6.97	3.58
✓	✓	✓	17.89	9.51	7.73	7.41	3.83

Table 5: Impact of different memory.

RPP	OE	MR	LMP	b+n	b+v	b+n+t	b+n+v	b+n+v+t
✓				17.35	8.39	6.93	5.92	2.99
✓	✓			17.17	8.75	7.13	6.52	3.48
✓	✓		✓	18.18	10.23	7.38	7.80	3.91
✓	✓	✓	✓	**19.12**	**11.06**	**8.31**	**8.09**	**4.43**

Table 6: Impact of different proposed components.

5.2 Implementation Details

For the first-stage object detection, we use Faster R-CNN pretrained on Ego4D. For the second stage, ViT-B serves as the backbone for video feature extraction. During training, only prediction boxes with IoU > 0.5 with the ground truth are utilized. Loss coefficients are balanced as $\lambda_1 = \lambda_2 = 10$ and $\tau = 3$. The teacher model employs the trained retentive encoder, while the student model uses the predictive encoder. We sample 8 frames with a stride of 8, resulting in $l = 64$. Implementation details of retentive and predictive pre-training are provided in the supplementary materials.

5.3 Comparison with the State of the Art

We evaluate EgoAnticipator against previous state-of-the-art models using mean Average Precision (mAP) metrics. In the STAv1 benchmark (Table 1), EgoAnticipator outperforms other two-stage methods [Chen *et al.*, 2022; Grauman *et al.*, 2022]. Notably, with a ViT-B backbone, EgoAnticipator exceeds InternVideo [Chen *et al.*, 2022], which utilizes a ViT-L pretrained on Ego4D, by +2.0 mAP. For the STAv2 dataset (Table 2), EgoAnticipator was evaluated by generating detection results in the first stage. Replicating the baseline with our detections and applying EgoAnticipator led to a significant mAP improvement of +2.7 over two-stage methods, nearing the performance of one-stage approaches. Furthermore, integrating our retentive and predictive encoders with the one-stage method StillFast [Tan *et al.*, 2023] further enhances performance. These results underscore the effectiveness of our retentive and predictive pre-training strategy. We also compare the performance of EgoAnticipator with $\phi^M = \phi^S$ and $\phi^M = \phi^T$. The findings indicate that the former achieves strong performance, while the latter can achieve further improvements by increasing computation during testing.

5.4 Ablation Studies

A detailed analysis of the EgoAnticipator was conducted on Ego4D-STAv1 using mAP as the main metric. A total of 4 ablation studies were conducted to assess the contributions of various components. *More detailed ablation studies can be found in the supplementary material.*

1) Retentive and predictive pre-training data. To evaluate RPP's effectiveness, we train multiple STA baseline models with video encoders pre-trained on different datasets, as detailed in Table 3. The results show that encoders trained on $\mathcal{D}^{\text{EgoIR}}$ outperform those pre-trained on K400 by +23.1%. Training on $\mathcal{D}^{\text{EgoIP}}$ yields an additional +4.6% improvement. Moreover, encoders trained on the complete Ego4D dataset, which includes approximately 0.6M exocentric action narrations, do not enhance the STA task, suggesting that exocentric information may not be beneficial.

2) Distillation supervision. In Table 4, we carry out diverse experiments to determine the supervision of distillation. Beyond KL divergence supervision for the classification of n and v, we explored time to contact δ, ranking score p, and feature supervision. The results show that as supervision increases, the student's performance continuously improves +7.0%, +4.7%, +0.3%, and +3.4%. These findings highlight the importance of comprehensive distillation supervision in bridging the gap between past and future representations. By learning from the teacher model's superior future insights, the student model aligns its feature distribution more closely with the teacher's, effectively reducing the performance disparity.

3) Impact of long-term memory. Table 5 shows how different memory types affect model performance. "short-term" is the baseline without long-term memory, "long-i" is our proposed long-term memory prompting, and "long-v" is the temporal memory method from LSTR [Xu *et al.*, 2021]. Temporal memory ("long-v") provides a modest improvement of +19.7% by capturing vague visual information. In contrast, long-term memory prompting with interactive memory embedding ("long-i") captures object-centric interactions, including verbs and contact times, resulting in a larger enhancement of +30.8%.

4) Overall ablation. As delineated in Table 6, we offer a comprehensive ablation study for each learning approach proposed within our research. The data, organized from top to bottom, reveals a sequential improvement in model performance. The contribution of Object Embedding (OE) is primarily in terms of mAP. Additionally, we observe that retentive and predictive pre-training (RPP), mirror distillation (MD), and long-term memory prompting (LMP) are instrumental in retaining memory and forecasting future events, thus significantly bolstering the STA task in terms of mAP.

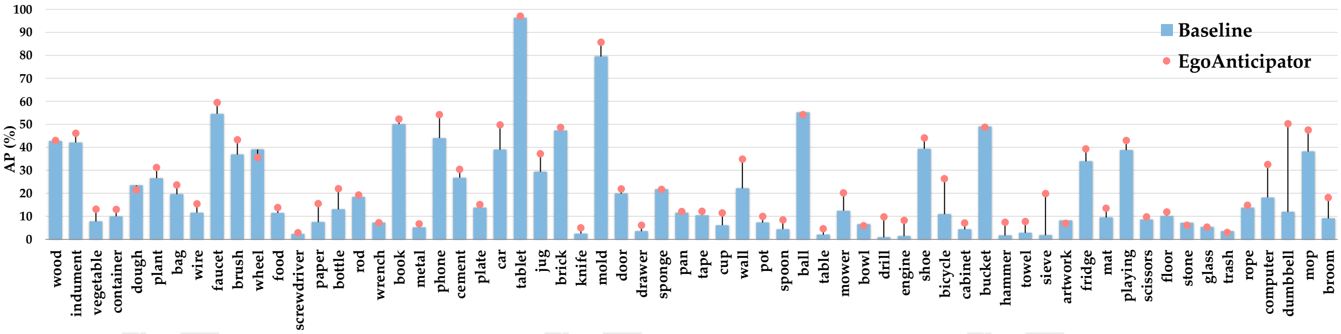

Figure 6: **Per-noun-category AP$_{b+n}$ on STAv1**: an improved baseline model with K400 pretrainning (15.52 mAP$_{b+n}$) *vs.* its EgoAnticipator counterpart (19.12 mAP$_{b+n}$). Categories are sorted by the number of examples.

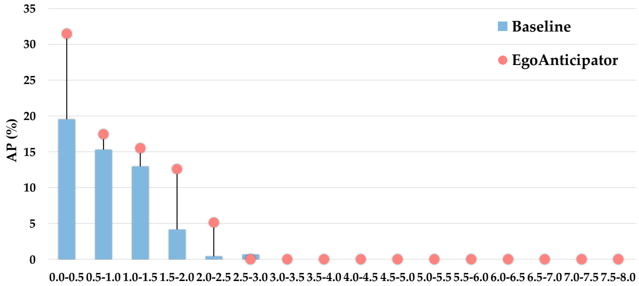

Figure 7: **Per-TTC-range AP$_{b+\delta}$ on STAv1**: an improved baseline model with K400 pretrainning (9.60 AP$_{b+\delta}$) *vs.* its EgoAnticipator counterpart (15.56 AP$_{b+\delta}$). We split the TTC into 16 bins, each spanning 0.5 seconds. The segmentation is based on the TTC statistics in the validation set.

5.5 Results Analysis

We present quantitative and qualitative results demonstrating EgoAnticipator's performance on STAv1.

Quantitative results. As shown in Figure 6, EgoAnticipator outperforms the baseline in 56 of 85 categories, notably with small objects like "bottle", "dumbbell", and "phone". TTC anticipation analysis (Figure 7) reveals significant improvements in short-term (0-0.5) and middle-term (1.5-2.5) predictions.

Qualitative results. Figure 8 displays two successful predictions and two failures. The first failure likely stems from sparse annotations, while the second indicates the model's difficulty distinguishing actions with similar motions but different semantics.

Efficiency analysis. Building on the Ego4D baseline, our framework utilizes ViT as the video encoder, offering a simpler and more efficient structure compared to SlowFast, especially with Flash Attention†. Table 7 presents relevant parameters and inference speeds on RTX 4090 GPU. Additionally, our decoder is streamlined with MLPs and a single attention layer, totaling only 13.5M parameters. When $\phi^M = \phi^S$, our model can achieve the best efficiency (121 FPS) and leading performance (4.29 mAP on STAv1). When $\phi^M = \phi^S$, our model can achieve the best performance (4.43 mAP on STAv1) yet relatively high FPS (59). Overall, EgoAnticipator is an efficient framework.

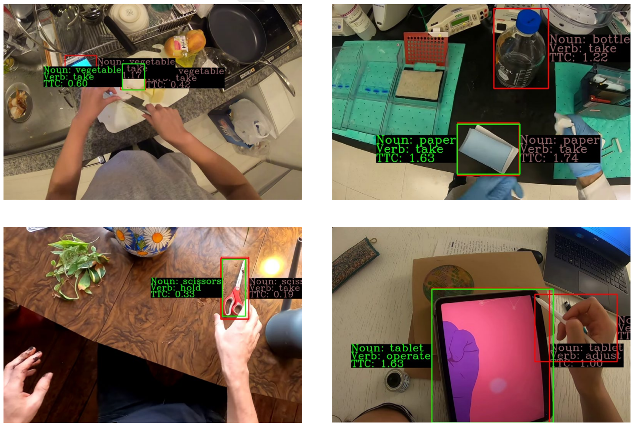

Figure 8: **Visualization on Ego4D STAv1**: Two success examples (top) and two failure cases (bottom).

Model	Param	FPS
SlowFast	34M	59
ViT-B	85M	128(161†)
ViT-B + EgoAnticipator ($\phi^M = \phi^S$)	98M	121(153†)
ViT-B + EgoAnticipator ($\phi^M = \phi^T$)	183M	59(74†)

Table 7: Efficiency comparison between EgoAnticipator and other video encoders on parameter (M) and inference speed (FPS).

6 Conclusion

In this paper, we introduce *EgoAnticipator*, a novel Retentive and Predictive Learning framework for egocentric object-interaction anticipation. EgoAnticipator improves anticipation capabilities by addressing misalignment in egocentric encoding, insufficient supervision signals, and suboptimal use of historical data. It incorporates long-term object-centric interactions and leverages both current video data and extended historical context for more accurate predictions. Experiments show that the retentive and predictive encoding mechanism, with mirror knowledge distillation, effectively shapes memory and introduces predictive uncertainty. Additionally, long-term prompting efficiently integrates historical information into the anticipatory process. EgoAnticipator unifies memory and predictive learning, achieving significant performance improvements for egocentric interaction anticipation.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (GrantNo. 62372223, U24A20330), in part by Nanjing University-China Mobile Communications Group Co., Ltd. Joint Institute (No. NR2310J7M), and the Fundamental Research Funds for the Central Universities.

Contribution Statement

Tong Lu served as the corresponding author for communication and correspondence.

References

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[Cai *et al.*, 2016] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016.

[Chen *et al.*, 2017] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *NIPS*, 30, 2017.

[Chen *et al.*, 2022] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *CoRR*, abs/2211.09529, 2022.

[Damen *et al.*, 2018] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, pages 720–736, 2018.

[Dong *et al.*, 2023] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey for in-context learning. *CoRR*, abs/2301.00234, 2023.

[Fan *et al.*, 2021] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021.

[Fernando and Herath, 2021] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13224–13233, 2021.

[Furnari and Farinella, 2020] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4021–4036, 2020.

[Furnari and Farinella, 2023] Antonino Furnari and Giovanni Maria Farinella. Streaming egocentric action anticipation: An evaluation scheme and approach. *Computer Vision and Image Understanding*, 234:103763, 2023.

[Furnari *et al.*, 2017] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *J. Visual Commun. Image Represent.*, 49:401–411, 2017.

[Gao and Wang, 2023] Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *ICCV*, pages 9901–9910, 2023.

[Girdhar and Grauman, 2021] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, pages 13505–13515, 2021.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[Grauman *et al.*, 2022] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.

[Hinton *et al.*, 2015] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[Huang *et al.*, 2018] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*, 2018.

[Huang *et al.*, 2022] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from A stronger teacher. In *NeurIPS*, 2022.

[Huang *et al.*, 2024] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, and Yu Qiao. Egoexolearn: A dataset for bridging asynchronous ego- and exocentric view of procedural activities in real world. In *CVPR*, pages 22072–22086, 2024.

[Kay *et al.*, 2017] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[Köpüklü *et al.*, 2019] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. *CoRR*, abs/1911.06644, 2019.

[Li *et al.*, 2021] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, pages 6943–6953, 2021.

[Lin *et al.*, 2022] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao,

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *NeurIPS*, 35:7575–7586, 2022.

[Liu *et al.*, 2019] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, pages 2604–2613, 2019.

[Min *et al.*, 2022] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, pages 11048–11064, 2022.

[Mullapudi *et al.*, 2019] Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *ICCV*, pages 3573–3582, 2019.

[Mur-Labadia *et al.*, 2024] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, José J. Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-ttention! affordances and attention models for short-term object interaction anticipation. In *ECCV*, pages 167–184, 2024.

[Murray *et al.*, 2012] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415, 2012.

[Núñez-Marcos *et al.*, 2022] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. *Neurocomputing*, 472:175–197, 2022.

[Patrick *et al.*, 2021] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *NeurIPS*, 34:12493–12506, 2021.

[Pirsiavash and Ramanan, 2012] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, pages 2847–2854, 2012.

[Ragusa *et al.*, 2023] Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *CVPR*, pages 3635–3644, 2023.

[Sener *et al.*, 2020] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *ECCV*, pages 154–171, 2020.

[Tan *et al.*, 2023] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. In *NeurIPS*, 2023.

[Thakur *et al.*, 2024] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Leveraging next-active objects for context-aware anticipation in egocentric videos. In *WACV*, pages 8642–8651, 2024.

[Tran *et al.*, 2021] Vinh Tran, Yang Wang, Zekun Zhang, and Minh Hoai. Knowledge distillation for human action anticipation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2518–2522. IEEE, 2021.

[Wang *et al.*, 2021] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *ICCV*, pages 8168–8177, 2021.

[Wang *et al.*, 2023] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *ICCV*, pages 20270–20281, 2023.

[Wu *et al.*, 2022] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *CVPR*, pages 13587–13597, 2022.

[Xu *et al.*, 2021] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *NeurIPS*, 34:1086–1099, 2021.

[Zhang and Ma, 2020] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*, 2020.

[Zhao and Krähenbühl, 2022] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *ECCV*, pages 485–502, 2022.

[Zhao *et al.*, 2020] Peisen Zhao, Jiajie Wang, Lingxi Xie, Ya Zhang, Yanfeng Wang, and Qi Tian. Privileged knowledge distillation for online action detection. *CoRR*, abs/2011.09158, 2020.