Interaction Region Visual Transformer for Egocentric Action Anticipation

Debaditya Roy[1], Ramanathan Rajendiran[1], and Basura Fernando[1,2]

[1]Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore
[2]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

Abstract

Human-object interaction (HOI) and temporal dynamics along the motion paths are the most important visual cues for egocentric action anticipation. Especially, interaction regions covering objects and the human hand reveal significant visual cues to predict future human actions. However, how to incorporate and capture these important visual cues in modern video Transformer architecture remains a challenge. We leverage the effective MotionFormer that models motion dynamics to incorporate interaction regions using spatial cross-attention and further infuse contextual information using trajectory cross-attention to obtain an interaction-centric video representation for action anticipation. We term our model InAViT which achieves state-of-the-art action anticipation performance on large-scale egocentric datasets EPICKTICHENS100 (EK100) and EGTEA Gaze+. On the EK100 evaluation server, InAViT is on top of the public leader board (at the time of submission) where it outperforms the second-best model by 3.3% on mean-top5 recall. The code is available[1].

1. Introduction

In egocentric action anticipation, the model needs to predict the immediate next human action that is going to happen, usually 1 second into the future [11]. Action anticipation is a challenging task due to various reasons such as the *uncertainty in future actions*, *diversity of execution of actions*, and the *complexity of human-object interactions* presented when executing those actions. A way to reduce the uncertainty in predicting the next action is to develop models that may be able to infer information about probable future objects and interactions that will be used in the future. As the observed and next actions are causally related, it is highly likely that information about observed interactions may possess possible cues about future actions.

[1]https://github.com/LAHAproject/InAViT

Some prior works use hand-object interactions to exploit those cues for action anticipation [48, 32], yet do not exploit the visual changes in human and object appearance caused by the execution of actions. We hypothesize that modeling of change in the appearance of regions containing objects and hands may reveal vital information about the probable execution of future actions. The work in [32] focuses on predicting manually annotated interaction hotspots without accounting for the specific objects associated with the interaction. However, it is challenging to automatically recognize the most important image regions that reveal what is going to happen next from egocentric videos. We develop an approach that attends to the most informative patches among the entire interaction region in every observed frame and make use of those modeled changes in appearances to predict future actions.

In particular, we propose to model the interaction region containing the hands and the interacted objects. We model the interaction regions as how hand and object appearance change due to the execution of the action and use those changes in appearance. Furthermore, we infuse the interaction regions with contextual cues from the background of the current action to obtain additional visual information about the area in which the interactions are performed. Finally, we propose an effective way to incorporate context-infused hand-object interaction regions to a video Transformer to create a richer interaction-centric video representation. Specifically, we use effective MotionFormer [43] as it aggregates important dynamic information along implicitly determined motion paths. This simple yet effective idea improves the action anticipation performance outperforming many dedicated video Transformer models.

Leveraging on the MotionFormer [43], we design a spatio-temporal visual transformer denoted as human-object Interaction visual transformer (InAViT) for action anticipation that refines image patches from the interaction regions in every frame which we term as *interaction tokens*. The interaction tokens are obtained from the refined object and human tokens and the refinement is influenced by objects, hands, and the anticipated action. We incorporate the visual context of the interaction's surroundings into the interaction tokens by proposing a trajectory cross-attention mechanism based on trajectory attention [43]. Finally, we infuse interaction tokens into the observed video to build an interaction-centric video representation for effective action anticipation. InAViT provides a way to extract visual changes in interaction regions across observed frames using trajectory cross-attention while in [48], human and object visual features are simply concatenated frame-wise to represent interactions.

Human-object interactions in actions are expressed as relations in spatio-temporal graphs in [24, 55, 36, 53, 42]. We are inspired by these approaches, especially the progression of human-object relationships for action recognition tasks [42]. We model interaction regions explicitly as spatio-temporal visual changes in hands and objects rather than implicitly as edges between hands and objects [42]. Furthermore, authors in [22, 60] make use of only object dynamics, and do not treat human hands as a separate entity [36, 42]. On the other hand, our interaction-centric video representation recognizes hands as a separate entity from objects that affect visual change on objects and vice-versa. The change in the hand's visual appearance when interacting with objects also gives a clue about the observed action. Hence, modeling interaction regions by capturing visual changes in both hands and objects is better than modeling the change in objects alone [22]. Therefore, we postulate that interaction-centric representations are better suited for action anticipation than object-centric approaches.

In summary, our contributions are twofold: (1) We propose a novel human-object interaction module that computes appearance changes in objects and hands due to the execution of the action and models these changes in appearances to refine the video representations of Video Transformer models for effective action anticipation. (2) On the EK100 evaluation server, our model **InAViT is top of the public leaderboard** and outperforms the second-best model by 3.3% on mean-top5 recall. Leveraging on MotionFormer, we also obtain massive improvement in EGTEA Gaze+ dataset.

2. Related Work

Anticipating human actions has gained interest in the research community with large datasets [19, 38, 6] and innovative approaches [11, 51, 17, 56, 9, 20, 57, 52, 47, 1, 44, 58, 26, 13, 12]. In [11], rolling and unrolling LSTM (RU-LSTM) is proposed to predict the next action. The authors in [51] increase RU-LSTM's temporal context using non-local blocks to combine local and global temporal context. In [17], spatio-temporal transformers called Anticipative Video Transformers (AVT) are proposed for action anticipation. In [56], MemViT extends AVT for long-range sequences by memory caching multiple smaller temporal sequences. RAFTformer [16] proposes a real-time action anticipation transformer using learnable anticipation tokens to capture global context trained using masking-based self-supervision. In [18, 40, 34, 41], present the problem of long-term anticipation (LTA). In [7], motion primitives called therbligs are introduced for decomposing actions which are then used for action anticipation. The authors in [49, 50] use goal representation for next action anticipation while [35] focuses on discovering intentions for LTA. The audio information is used to augment video representation for next-action anticipation in [62] and [39] uses only audio for LTA. Other approaches consider past and future correlation using Jaccard vector similarity [9], self-regulated learning [44], transitional model [37], and counterfactual reasoning [61].

There are a variety of approaches for Human-Object Interactions (HOIs) modeling in images [4, 15, 30, 14, 59, 5, 27, 28, 31] and videos [32, 29, 33, 25, 8]. In [29], human-object interaction regions are detected in videos using verb-object queries describing the action. Observed action labels are not available during testing in action anticipation and hence, we cannot predict the interaction region. In [25], relationships between humans and objects are modeled and verified using relationship labels. Relationship labels are not available for the observed video and so our model attends to all interactions and discovers their importance in predicting the next action.

Another related area is interaction hotspot prediction where future hand-trajectory and interaction spots need to be estimated. Interaction prediction approaches [32, 33] learn future hand motion distribution conditioned on the video representation using an encoder (LSTM [32] or Transformer [33]). Object interaction anticipation [45] requires predicting the next active object bounding box along with the next action. However, these approaches require explicit annotations of hand trajectories in future, object trajectories in the observed frames and location of interaction spots. Our interaction modeling approach obviates the need for hand and object trajectory annotations that are difficult to obtain in videos as shown in [10].

3. Preliminaries

3.1. Basic video representation

We extract a set of 3-D cuboids or video "tokens" from a video as existing video transformers [2, 43, 3]. The set of all the video tokens from a single fixed length video-clip is denoted by $X \in \mathbb{R}^{THW \times d}$ wherein each of the THW cuboids is linearly projected to a d-dimensional vector. Here, T is the number of frames in the fixed-length video clip and H, W is the number of vertical and horizontal patches respectively. Let $\mathbf{x}_{st} \in \mathbb{R}^d$ denote a video token from the set X at spatial location $s \in \{1, \cdots, H \times W\}$ and temporal location $t \in \{1, \cdots, T\}$. Similar to [43], we add separate learnable positional encoding for spatial and temporal dimension for each video token denoted as $\mathbf{e}_s^s \in \mathbb{R}^d$ and $\mathbf{e}_t^t \in \mathbb{R}^d$, respectively. The resultant video token after spatial and temporal embedding is given as $\mathbf{x}_{st} = \mathbf{x}_{st} + \mathbf{e}_s^s + \mathbf{e}_t^t$. A classification token $\mathbf{x}_{cls}$ is appended for anticipating the next action from X resulting in $THW+1$ tokens in $\mathbb{R}^d$. We exclude the classification token hereafter for clarity.

3.2. Obtaining hand and object tokens

We obtain hand and object representations from video tokens X. We obtain object and hand bounding-boxes using Faster R-CNN [46]. In every frame, we use one bounding box for the hand and N bounding boxes for objects closest to the hand. SORT algorithm is used over the detections to obtain sequences of detections where each sequence represents a hand or an object [42]. Now, given the video tokens corresponding to a frame X_t and the bounding box of the hand $B_{h,t}$, we obtain a hand token $\mathbf{h}_t \in \mathbb{R}^d$. We make use of RoIAlign [21] layer on X_t to obtain hand region crops similar to [22]. We then use MLP and max-pooling to obtain the final hand representation or the hand token. We apply this to every frame to obtain T hand tokens denoted by $\mathcal{H} \in \mathbb{R}^{T \times d}$ where $\mathcal{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_T]$. Similarly, for each of object i, we obtain a T object tokens $\mathbf{o}_{i,1}, \cdots, \mathbf{o}_{i,T}$ and we denote it as $\mathcal{O}_i \in \mathbb{R}^{T \times d}$. In total, for the N objects, we end up with $\mathcal{O} \in \mathbb{R}^{T \times N \times d}$ where $\mathcal{O} = [\mathcal{O}_1, \cdots, \mathcal{O}_N]$.

4. Our method

4.1. Overview of our method

We hypothesize that in egocentric action anticipation, hands and objects play a key role in anticipating actions. Hands and objects change the appearance of other objects causing visible state changes, such as *human cutting tomato with knife* or *human emptying a pan using spatula*. Change in the state of the objects reveals cues about the possible next action. We capture these changes using newly designed *interaction tokens* by refining the original hand and object representations (tokens) with respect to each other.

As the objects affect the appearance of hand regions, we refine the hand tokens using all object tokens in the frame using Eq. (1). Similarly, as hand and objects affect the appearance of other objects when executing the action, we model this by refining every object token using hand tokens and other object tokens in the frame using Eq. (2).

$$\tilde{\mathcal{H}} = \phi_{\mathcal{H}}(\mathcal{H}|\mathcal{O}) \tag{1}$$

$$\tilde{\mathcal{O}}_i = \phi_{\mathcal{O}}(\mathcal{O}_i|\mathcal{H}, \mathcal{O}_j)\ \forall j \neq i, i \in 1, \cdots, N, \tag{2}$$

Here $\phi_{\mathcal{H}}$ and $\phi_{\mathcal{O}}$ are attention-based functions that will be discussed in detail in Secs. 4.2.1 to 4.2.3. Refined object tokens for all objects are denoted as $\tilde{\mathcal{O}} = [\tilde{\mathcal{O}}_1, \cdots, \tilde{\mathcal{O}}_N]$. Together, the refined hand and object tokens constitute the interaction tokens $I = [\tilde{\mathcal{H}}, \tilde{\mathcal{O}}]$.

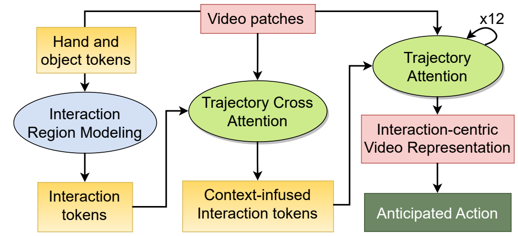


Figure 1: Block diagram of our interaction region modeling based action anticipation (InAVIT). Ellipses represent processing mechanisms. Trajectory attention $\times 12$ refers to MotionFormer.

We also hypothesize that the context or the background of the current action may provide useful information when predicting the next action. For example, *picking a tomato* next to the cutting board (context) informs that the next probable action is *cut tomato*. Therefore, we enrich the interaction tokens (I) using the information coming from the context and vice-versa. We use Trajectory Cross Attention function (ϕ_I) inspired by [43] to obtain the context-infused interaction tokens $\tilde{I}$

$$\tilde{I} = \phi_I(I|X). \tag{3}$$

The final video representation is interaction-centric where we first concatenate context-infused interaction tokens to the video tokens. Then, we assimilate the interaction regions into the video using self-attention (ϕ_X). We choose the refined tokens corresponding to the original video tokens$X$ as the interaction-centric video representation X_I

$$X_I = \phi_X([\tilde{I}, X]). \tag{4}$$

We predict the next action using the interaction-centric video representation with multiple layers of Trajectory Attention as in MotionFormer [43]

$$\mathbf{a}_{next} = \phi(X_I). \tag{5}$$

Next action anticipation is defined as observing $1, \cdots, T$ frames and predicting the action that happens after a gap of T_a seconds. It is important to note that a new action starts after T_a seconds that is not seen in the observed frames. Our overall approach is shown in Fig. 1.

4.2. Human-Object interaction region modeling

Now we discuss how we implement Eq. (1) and Eq. (2). We model spatiotemporal interaction regions between hands and objects in three ways encapsulating different types of interaction information to obtain *interaction tokens* $I = [\tilde{\mathcal{H}}, \tilde{\mathcal{O}}]$. These three types of interaction modeling

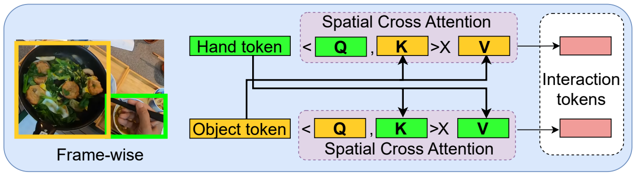


Figure 2: Modeling interaction region tokens using Spatial Cross Attention. In every frame, hand tokens act as query and object tokens as key and value to compute refined hand tokens. Refined object tokens are computed with object token as query, and hand and other object tokens as key and values (not shown here to avoid clutter). Interaction tokens consist of refined hand and object tokens.

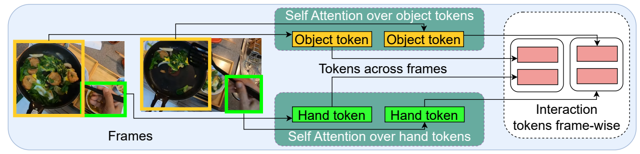


Figure 3: Modeling interaction tokens using Self-attention Over Time. We compute self-attention over hand tokens in all the frames to obtain refined hand tokens. We repeat this for every object region across frames. The refined hand and object tokens at every frame are the interaction tokens.

will be evaluated in the experiments. Reader may refer to the supplementary material section 1 for the common definition of cross-attention and self-attention. Next, we present three ways to obtain interaction tokens.

4.2.1 SCA: Spatial cross-attention

We implement Eq. (1) and Eq. (2) using spatial cross-attention. In every observed frame, there is a hand token and multiple object tokens. We model the change in hands by cross-attention [54] as shown in Fig. 2. We use the hand token $\mathbf{h}_t$ as the query and compute the attention with respect to every object token in the same frame $\mathbf{o}_{1,t}, \cdots, \mathbf{o}_{N,t}$. The query, key, and value are denoted as $\mathbf{q}_{h,t} = \mathbf{h}_t \mathbf{W}_q$, $\mathbf{k}_{i,t} = \mathbf{o}_{i,t} \mathbf{W}_k$, and $\mathbf{v}_{i,t} = \mathbf{o}_{i,t} \mathbf{W}_v$, respectively. We use cross-attention to get the refined hand tokens, $\tilde{\mathbf{h}}_t$. Cross attention implicitly seeks the object token that has the most impact on the hand token by pooling all the object tokens in the frame and weighing each by its probability.

Similarly, we use cross-attention to refine each object token using the hand token and other object tokens in the frame. Every object token $\mathbf{o}_{i,t}$ acts as a query, and human and other object tokens act as keys and values. Let $\mathbf{z}_t$ represent either hand or other object tokens in the frame t. There are N such tokens in each frame for each object query $\mathbf{o}_{i,t}$. The query, key, and values are obtained as $\mathbf{q}_{i,t} = \mathbf{o}_{i,t} \mathbf{W}_q$, $\mathbf{k}_{j,t} = \mathbf{z}_t \mathbf{W}_k$, and $\mathbf{v}_{j,t} = \mathbf{z}_t \mathbf{W}_v$, respectively. The refined object token $\tilde{\mathbf{o}}_{i,t}$ are obtained using cross-attention. We call the refined hand and object tokens as interaction tokens $I_t = [\tilde{\mathbf{h}}_t, \tilde{\mathbf{o}}_{1,t}, \cdots, \tilde{\mathbf{o}}_{N,t}]$. We perform spatial cross-attention (SCA) over every frame to obtain all the interaction tokens $I \in \mathbb{R}^{T \times (N+1) \times d}$.

4.2.2 SOT: Self-attention of hand/object over time

We model interaction tokens as the change in hands or objects individually over time as shown in Fig. 3 using self-attention. Hand token h_t is refined using only other hand tokens from all frames. The query, key, and value are obtained from all hand tokens across all frames $\mathbf{q}_{h,t} = \mathbf{h}_t \mathbf{W}_q$, $\mathbf{k}_{h,t} = \mathbf{h}_t \mathbf{W}_k$, $\mathbf{v}_{h,t} = \mathbf{h}_t \mathbf{W}_v$ to obtain refined hand token $\tilde{\mathbf{h}}_t$ using self-attention. We refine object tokens $\mathbf{o}_{i,t}$ of every object i separately over frames using self-attention. The query, key, and value for object i are computed from its own tokens over all the frames $\mathbf{q}_{i,t} = \mathbf{o}_{i,t} \mathbf{W}_q$, $\mathbf{k}_{i,t} = \mathbf{o}_{i,t} \mathbf{W}_k$, $\mathbf{v}_t = \mathbf{o}_{i,t} \mathbf{W}_v$ to obtain refined object token $\tilde{\mathbf{o}}_{i,t}$ using self-attention. We call this method Self-attention Over time **SOT** and SOT interaction tokens $I_t = [\tilde{\mathbf{h}}_t, \tilde{\mathbf{o}}_{1,t}, \cdots, \tilde{\mathbf{o}}_{N,t}]$ consist of refined hand and object tokens.

4.2.3 UB: Union Box of hand and nearest object

We obtain the third type of interaction token using the hand and the nearest object in every frame. We compute the union bounding box from the hand and the nearest object bounding boxes in every frame. Here the hand and object union box is similar to the union of objects and human regions for interaction detection [29]. We then obtain the union tokens $\mathcal{U} \in \mathbb{R}^{T \times d}$ using the method described in Sec. 3.2. Unlike the previous two approaches, the union region consists of both human and object features together. We refine the T union tokens in $\mathcal{U} \in \mathbb{R}^{T \times d}$ using self-attention to obtain interaction tokens $I_t \in \mathbb{R}^{T \times d}$. We denote this approach as *UB* for short. Next, we describe how to obtain context-infused interaction tokens in Sec. 4.3. Then, in Sec. 4.4, we describe how context-infused interactions are used to obtain interaction-centric video representation for action anticipation.

4.3. CI: Context-infused interaction tokens

Here we discuss how we implement Eq. (3). The context plays an important role along with interaction in deciding what are the possible next actions. For example, *washing a plate* in *kitchen sink* suggests that the next action is probably *close the tap*. Hence, we infuse context into interaction tokens by proposing Trajectory Cross Attention (TCA) based on trajectory attention [43]. TCA maintains temporal correspondences between interaction tokens and context tokens of a frame as shown in Fig. 4.

Our TCA formulation seeks the probabilistic path of an interaction token between frames. The interaction tokens act as the query on the video tokens X that is representative of the context. Let the video patch at spatial location s in frame t be given by $\mathbf{x}_{st} \in \mathbb{R}^d$. The key and value are obtained from the video patch. We have $N+1$ interaction tokens[2] in every frame. For each interaction token $\mathbf{y}_t \in I_t$, we obtain a set of trajectory tokens $\hat{\mathbf{y}}_{tt'} \in \mathbb{R}^d, \forall t' \geq t$ that represents pooled information weighted by the trajectory probability. The pooling operation implicitly looks for the best location s at frame $t' \geq t$ by comparing the interaction query $\mathbf{q}_t = \mathbf{y}_t \mathbf{W}_q$ to the context keys $\mathbf{k}_{st'} = \mathbf{x}_{st'} \mathbf{W}_k$ using q, k, v-attention. Attention is applied spatially and independently for all the interaction tokens in every frame. This is complementary to our previously computed cross attention (Eq. (1) and Eq. (2)) where hand/object query tokens are refined with respect to other object/hand key and value token. In TCA, we seek to infuse interaction tokens with the visual context of the video which is similar to [23] where cross-attention is used to infuse query tokens with information from key and value tokens.

Once trajectories are computed, we pool them across time to reason about connections across the interaction regions in a frame given the environment. For temporal pooling, the trajectory tokens are projected to a new set of queries, keys, and values $\hat{\mathbf{q}}_{tt'} = \hat{\mathbf{y}}_{tt'} \mathbf{W}_q$, $\hat{\mathbf{k}}_{tt'} = \hat{\mathbf{y}}_{tt'} \mathbf{W}_k$, $\hat{\mathbf{v}}_{tt'} = \hat{\mathbf{y}}_{tt'} \mathbf{W}_v$, respectively. The new query $\hat{\mathbf{q}}_{tt'}$ has information across the entire trajectory that extends across the entire observed video frames. We perform temporal pooling using 1D attention across the new time (trajectory) dimension to obtain refined interaction tokens. We term these refined interaction tokens as *context-infused interaction tokens* $\tilde{I} = [\tilde{\mathbf{y}}_1, \cdots, \tilde{\mathbf{y}}_T]$.

4.4. ICV: Interaction-centric Video Representation

Here we discuss how we implement Eq. (4). The next action is dependent on interactions and context. Therefore, our goal is to obtain a video representation that incorporates information from the context-infused interaction tokens. We concatenate the video tokens X and the context-infused interaction tokens $\tilde{I}$ to form the augmented video representation X_a[3]. Then we perform self-attention over the augmented video tokens $\mathbf{x}^a_r \in X_a$ wherein video tokens attend over the interaction regions tokens and vice-versa. After self-attention, we obtain refined augmented video tokens $\tilde{\mathbf{x}}^a_r$. We construct the interaction-centric video representation X_I using $\tilde{\mathbf{x}}^a_r$ that corresponds to the original video tokens.

$$X_I = \tilde{\mathbf{x}}^a_r, \forall r, \text{ if } \mathbf{x}^a_r \in X. \tag{6}$$

We call X_I as interaction-centric video representation in the same vein as object-centric video representation [22].

Till now, we have only applied a single attention layer for interaction modeling, context infusion for interaction, and interaction-centric video representation. In literature, video transformer approaches [43, 42, 2] have shown excellent performance with multiple layers of attention. We apply 12 layers of Trajectory Attention following MotionFormer [43] on the interaction-centric video representation to obtain the final video representation $\tilde{X}_I$. The reason for choosing MotionFormer is that it performs best empirically. The classification token $\mathbf{x}_{cls}$ obtained at the end of multiple trajectory attention layers is used to predict the next action, i.e., $\hat{\mathbf{a}}_{next} = \phi(\tilde{\mathbf{x}}_{cls})$ where ϕ is a linear layer.

Loss function. We use cross-entropy loss for the next action label to train our model. We compare the model's prediction $\hat{\mathbf{a}}_{next}$ with the ground truth one-hot label $\mathbf{a}_{next}$ for the next action as follows

$$\mathcal{L}_{ant} = -\sum \mathbf{a}_{next} \odot \log(\hat{\mathbf{a}}_{next}). \tag{7}$$

It should be noted that as we train our model with the cross-entropy loss to predict the next action, the interaction tokens are optimized for finding the most influential interactions when predicting the next action (see Fig. 5).

5. Experiments and Results

5.1. Datasets and Implementation Details

We evaluate and compare our methods on two large unscripted action anticipation datasets *EPIC-*

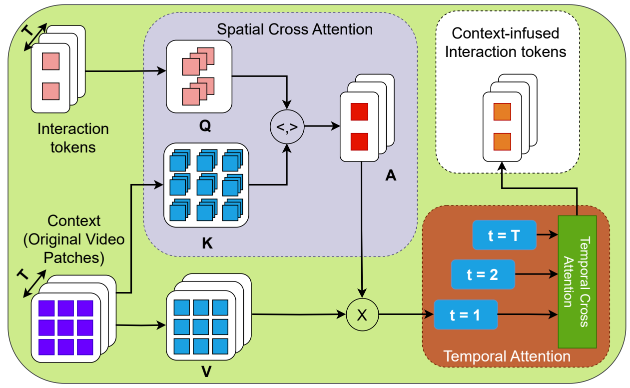


Figure 4: Context infusion into interaction region tokens using Trajectory Cross Attention. We compute spatial cross-attention (SCA) to find the best location for interaction trajectory by comparing the interaction query to context keys. Next, we pool the interaction trajectories across time to form connections across the interaction tokens in a frame.

[2] 1 in case of UB interaction modeling

[3] $X_a \in \mathbb{R}^{T\times(H\times W+N+1)\times d}$ (for SCA and SOT) or $X_a \in \mathbb{R}^{T\times(H\times W+1)\times d}$ for UB

KITCHENS100[6] (EK100) and *EGTEA Gaze+*[38]. For EK100, we report results on the test set evaluation server that uses mean-top5 recall as the metric. For EK100 and EGTEA, anticipation gap (T_a) is 1s and 0.5s, respectively.

We follow MotionFormer [43] and use 16-frame long clips of resolution 224×224 uniformly sampled from an observed video of 64 frames (approximately 2s). Every 3D video token is extracted from a video patch of size $2 \times 16 \times 16$. We extract hand, object, and union tokens following the strategy explained in Sec. 3.2. Then, to implement InAViT(SCA) (Sec. 4.2.1), we use a single cross-attention layer with 12 attention heads. Similarly, InAViT(SOT) and InAViT(UB) are implemented with a self-attention layer with 12 heads. We set the number of objects per frame as 4 for EK100 and 2 for EGTEA based on empirical performance (this also makes sure batch processing is efficient). We report the results of varying the number of objects per frame in Supplementary material section 3. If there are fewer objects (less than 4 or 2 respectively), then we zero pad them using null tokens and mask them, which will not impact the model. The same configuration of objects is used for training other baseline models such as ORViT-MF [22].

EK100 provided hand detections do not contain hand annotations for 20% of the frames. Since our aim is not localize hands accurately, we reduce the threshold to 0.05 from 0.1 used by EK100 to get hand detections in all frames. Lowering the threshold introduces bounding-boxes that cover the region around the hand that is useful for interaction region modeling and anticipation as shown in Fig. 6. The number of hand regions per-frame is set to 1 as both hands are not visible in most frames. If there are two hands in a frame, we randomly pick one and track it using SORT. Some analysis and statistics about detected objects in frames are shown in Supplementary material section 2.

We use one layer of trajectory cross-attention with 12 attention heads and a temporal resolution of 8 to obtain the context infusion of interaction tokens (Sec. 4.3). We then concatenate the refined interaction and original video tokens to form the augmented video tokens. We use a single self-attention layer with 12 heads on the augmented video tokens to obtain interaction-centric video tokens (Sec. 4.4). Finally, we apply MotionFormer on the interaction-centric video tokens to predict the next action (Sec. 4.4). We use a batch size of 16 video(clips) to train on 4 RTX A5000 GPUs with 24 GB memory each and the learning rate is set to $1e^{-4}$ with AdamW optimizer. We will release our code.

5.2. Ablation on InAViT

Component-wise validation. In Tab. 1(a), we show the contribution of each component of InAViT - interaction modeling using SCA (Eqs. (1) and (2)), **C**ontext **I**nfusion of interaction tokens (**CI**) (Eq. (3)), and **I**nteraction-

Method	Overall Action(%)	Unseen Action(%)	Tail Action(%)
SCA	12.66	15.49	06.03
SCA + CI	14.21	14.26	09.12
SCA+ICV	22.21	20.85	17.07
SCA + CI + ICV	**23.75**	**23.49**	**18.11**
(a) Component-wise validation of InAViT			
UB+CI + ICV	22.75	22.14	17.04
SOT+CI + ICV	22.48	20.56	17.46
(b) Comparing interaction modeling methods			
SCA-(Hand) + CI + ICV	23.27	23.21	17.57
SCA-(Obj) + CI + ICV	22.49	22.23	16.73
(c) Comparing refined hand vs. object as interaction tokens			

Table 1: Ablation of InaViT on EK100 evaluation server [**Test set**]. Verb and Noun results are in Supplementary.

Centric **V**ideo representation using trajectory attention (**ICV**) (Eq. (4)). Context infusion improves overall performance but we see the biggest improvement when we add the interaction-centric video representation (i.e. Eq. (4)). So, we conclude that interactions and the interaction centric video representations are important for action anticipation.

Comparing different interaction regions. In Tab. 1(b), we compare the three models for interaction region modeling - spatial cross attention (SCA+ CI + ICV), spatial attention over time (SOT+ CI + ICV), and union boxes (UB + CI + ICV) described in Sec. 4. In our comparison, SCA performs the best on overall, unseen, and tail classes compared to both SOT and UB. SCA contextualizes the visual change of each hand/object better using other objects compared to SOT which computes visual change individually. UB's focus is narrower than SCA as it only considers the nearest object and potentially leaves out other objects that can be used in the next action. SCA performs much better in tail classes where few examples are available and the model relies on visual information for anticipation. As SCA uses all the objects to model interactions, it extracts the most visual information from every frame to make better predictions.

Only Hand/Object as interaction regions. The best-performing SCA interaction model involves both refined hand and refined object tokens as interaction region tokens. In Tab. 1(c), we compare the contribution of only refined hand (SCA(Hand)+CI + ICV) and only object tokens (SCA(Obj)+CI + ICV) by using either of them as interaction region tokens. As in SCA formulation, we refine hand tokens with objects and object tokens with hand and other objects. Refined hand tokens perform better than refined object tokens as the position of the hand is vital in determining what object(s) can be used next. Still, interaction region tokens containing both refined hands and object tokens (SCA+ CI + ICV) perform the best. We conclude that modeling the changes in hand and object tokens provides useful information to improve action anticipation.

Effect of infusing context. We evaluate the effect of context infusion on interaction tokens in Tab. 2. For this comparison, we change either the input or mecha-

Method	Overall Action(%)	Unseen Action(%)	Tail Action(%)
SCA + CI + ICV	**23.75**	**23.49**	**18.11**
SCA + CI(Mask FG) + ICV	08.05	05.92	05.84
SCA+ Concat + ICV	22.14	23.47	17.24

Table 2: Effect of infusing context in different ways. CI(Mask FG): Context Infusion with foreground hands and objects masked out, Concat: Context infusion by concatenating context tokens with interaction tokens. Results are on evaluation server [**Test set**].

nism of context infusion(Sec. 4.3). Interaction modeling is done using SCA and interaction-centric video representation is computed using the original video tokens. For CI (Mask FG), we mask the foreground i.e., hands and objects from the context. We crop the objects and hands based on the bounding boxes and apply a Gaussian filter to soften the edges. We call this masked foreground context and use it to refine interaction tokens. The performance of (SCA+CI(Mask FG)+ICV) is quite poor which means that the complete context with foreground objects (SCA+CI+ICV) is better for refinement. We also find that the concatenation of context (video tokens) to interaction tokens (SCA+Concat+ICV) is worse than our proposed (SCA+CI+ICV) approach.

5.3. Comparison with state-of-the-art

We now compare our best-performing InAViT (SCA+CI+ICV) model against state-of-the-art approaches. On EK100 evaluation server, InAViT significantly outperforms other approaches as seen in Tab. 3. InAViT's performance is much better than AVT [17], MeMViT [56], and RAFTformer [16] on EK100 which also use visual transformers for representing the video. We also compare InAViT against the baseline MotionFormer (MF) [43] and object-centric video representation ORViT-MF [22]. As ORViT and MF are not trained for action anticipation, we train them using the official repositories.

We are the first to show the effectiveness of MF and ORVIT-MF for action anticipation which alone outperforms prior state-of-the-art methods. Our approach InAViT performs even better than both MF and ORVIT-MF and achieves significantly better results than the previous best results, the Abstract Goal [50] on EGTEA (Tab. 4) and AVT [17] on EK100. In fact, InAViT outperforms [50] by 18% in mean accuracy and 20.8% in top-1 accuracy on EGTEA. It also outperforms the human-object interaction method in [32] by 30% and 35% on mean and top-1 accuracy, respectively. Similarly, InAViT outperforms AVT [17] by 22%, 17%, and 7%, in the overall verb, noun, and action anticipation on EK100 which is impressive given the large number of actions (3805), nouns (300), and verbs (97). On EK100, InAViT outperforms AVT on unseen and tail action anticipation by 12% and 4%, respectively. This demonstrates that InAViT is effective at predicting rare actions and generalizes much better to new environments.

Set	Method	Overall (%)			Unseen (%)			Tail (%)		
		Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
Val	RU-LSTM [6]	23.20	31.40	14.70	28.00	26.20	14.50	14.50	22.50	11.80
	T. Agg. [51]	27.80	30.80	14.00	28.80	27.20	14.20	19.80	22.00	11.10
	AFFT [62]	22.80	34.60	18.50	24.80	26.40	15.50	15.00	27.70	16.20
	AVT [17]	28.20	32.00	15.90	29.50	23.90	11.90	21.10	25.80	14.10
	TrAc [20]	35.04	35.49	16.60	34.64	27.26	13.83	30.08	33.64	15.53
	MeMViT [56]	32.20	37.00	17.70	28.60	27.40	15.20	25.30	31.00	15.50
	Rformer [16]	33.80	37.90	19.10	-	-	-	-	-	-
	MF*	47.14	46.92	21.52	42.33	47.22	21.99	40.47	38.66	16.36
	ORViT-MF*	46.71	47.91	23.54	38.99	45.32	23.60	37.28	37.81	17.10
	InAViT (Ours)	**52.54**	**51.93**	**25.89**	**46.45**	**51.30**	**25.33**	**45.34**	**39.21**	**20.22**
Test	RU-LSTM [6]	25.25	26.69	11.19	19.36	26.87	09.65	17.56	15.97	07.92
	T. Agg. [51]	21.76	30.59	12.55	17.86	27.04	10.46	13.59	20.62	08.85
	AFFT [62]	20.70	31.80	14.90	16.20	27.70	12.10	13.40	23.80	11.80
	AVT [17]	26.69	32.33	16.74	21.03	27.64	12.89	19.28	24.03	13.81
	Abs. goal [50]	31.40	30.10	14.29	31.36	35.56	17.34	22.90	16.42	07.70
	TrAct [20]	36.15	32.20	13.39	27.60	24.24	10.05	32.06	29.87	11.88
	Rformer [16]	30.10	34.10	15.40	-	-	-	-	-	-
	MF*	45.14	45.97	19.75	40.36	45.28	19.49	39.17	35.91	14.11
	ORViT-MF*	43.74	46.61	21.53	38.99	45.32	21.47	37.28	35.78	15.96
	InAViT (Ours)	**49.14**	**49.97**	**23.75**	**44.36**	**49.28**	**23.49**	**43.17**	**39.91**	**18.11**

Table 3: Comparison with state-of-the-art on EK100 validation (Val) and evaluation server (Test). InAViT significantly outperforms other Transformer-based approaches such as AVT, MotionFormer and object-centric MotionFormer (ORViT-MF). *We trained MF and ORViT-MF for action anticipation using their official repositories.

Method	Top-1 Accuracy (%)			Mean Class Accuracy (%)		
	VERB	NOUN	ACT.	VERB	NOUN	ACT.
FHOI [32]	49.0	45.5	36.6	32.5	32.7	25.3
AFFT [62]	53.4	50.4	42.5	42.4	44.5	35.2
AVT [17]	54.9	52.2	43.0	49.9	48.3	35.2
Abs. Goal [50]	64.8	65.3	49.8	63.4	55.6	37.4
MF*	77.8	75.6	66.6	77.5	72.1	56.9
ORVIT-MF*	78.8	76.3	67.3	78.8	75.8	57.2
InAViT (Ours)	**79.3**	**77.6**	**67.8**	**79.2**	**76.9**	**58.2**

Table 4: Comparison of anticipation performance on EGTEA Gaze+. Complete table in Supplementary.

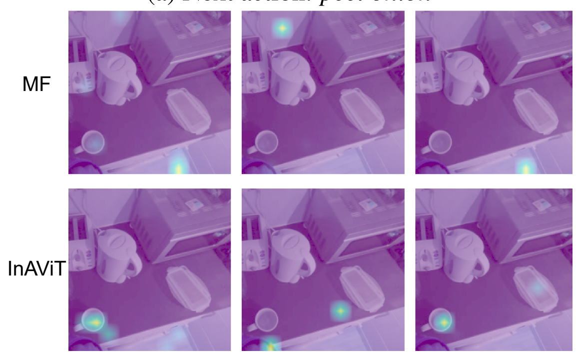
Top-1 Action Acc.	Anticipation gap (T_a in s)				GFlops	Params
	2	1.5	1	0.5		
MF	60.2	61.7	64.2	66.6	370	143.9M
ORVIT-MF	62.2	63.6	66.1	67.3	403	172.1M
InAViT	**64.1**	**65.8**	**66.9**	**67.8**	391	157.2M

Table 5: InAViT performs better even on larger anticipation gap of 1.5 and 2s (EGTEA). InAViT has better performance than ORVIT with lower computations and parameters.

In Tab. 5, we vary the anticipation gap (T_a) for EGTEA. InAViT performs better in longer anticipation as it uses human-object interaction information. We analyzed the computational cost and number of parameters of InAViT. The baseline MotionFormer requires 370 GFlops with 143.9M parameters, while InAViT requires 391 GFlops with 157.2M parameters and ORVIT-MF requires 403 GFlops with 172.1M parameters. InAViT is less computationally expensive and has a lower number of parameters than ORVIT-MF.

Qualitative Results. In Fig. 5, we visualize the attention

(a) Next action: *peel onion*



(b) Next action: *pour sugar*

Figure 5: InAViT attends to the location(s) where the next action will occur in (a) onion and (b) cup and sugar.

(a) GT: *fill kettle*, Pred: *fill kettle*

(b) GT: *wash glass*, Pred: *wash glass*

(c) GT: *drink water*, Pred: *drink water*

(d) GT: *adjust hob*, Pred: *adjust hob*

Figure 6: Some examples of InAViT's anticipation where green is correct. InaViT can anticipate correctly with hand detections (in blue) that precisely capture the hand (a)(b) or cover the region around the hand in (c)(d).

map of the $\mathbf{x}_{cls}$ token used for action anticipation on all spatial tokens across the frames. This helps us understand where InAViT focuses compared to MotionFormer. Motionformer attention is divided into many areas while InAViT attends to the interaction. While anticipating *peel onion* (Fig. 5(a)), InAViT pays high attention to the exact location the onion is being peeled. Similarly, when anticipating *pour sugar* (Fig. 5(b)), InAViT attends to both the cup and the sugar container. The frames for visualization are chosen based on the significant motion of hands and objects during the observed action. In Fig. 6(c)(d), we show that InAViT anticipates the action correctly even if the bounding box covers the region around the hand. We show more qualitative results in Supplementary.

6. Discussions and Conclusion

We present an effective method to improve ego-centric action anticipation by capturing human-object interaction information using a Transformer architecture. We showed that our spatial cross-attention (SCA) based human-object interaction information extraction along with the trajectory attention-based context infusion (CI) and the interaction-centric video (ICV) representations are effective in ego-centric action anticipation. It is interesting to note that our model obtains a 6.0% improvement over published results on EK100 test dataset and a massive 20.0+% improvement on EGTEA Gaze+ dataset. However, the biggest improvement comes from trajectory attention-based MotionFormer. We improve MotionFormer by 4.0% and 1.3% on EK100 and EGTEA Gaze+ datasets, respectively. The findings of this work advance action anticipation research.

Acknowledgment This research/project is supported in part by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award Number: AISG-RP-2019-010) and by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-016). This research is also supported by funding allocation to B.F. by the Agency for Science, Technology and Research (A*STAR) under its SERC Central Research Fund (CRF), as well as its Centre for Frontier AI Research (CFAR).

References

[1] Mohammad Sadegh Aliakbarian, Fatemehsadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *IEEE International Conference on Computer Vision ICCV 2017*, 2017. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2, 5

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 2

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2

[5] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 2

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. 2, 6, 7

[7] Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Therbligs in action: Video understanding through motion primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10626, 2023. 2

[8] Gueter Josmy Faure, Min-Hung Chen, and Shang-Hong Lai. Holistic interaction transformer network for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3340–3350, 2023. 2

[9] 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. 2

[10] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4991–5000, 2018. 2

[11] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 1, 2

[12] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Forecasting future action sequences with neural memory networks. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 298. BMVA Press, 2019. 2

[13] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5562–5571, 2019. 2

[14] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. 2

[15] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2

[16] Harshayu Girase, Nakul Agarwal, Chiho Choi, and Karttikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18759–18769, 2023. 2, 7

[17] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 2, 7

[18] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022. 2

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[20] Xiao Gu, Jianing Qiu, Yao Guo, Benny Lo, and Guang-Zhong Yang. Transaction: ICL-SJTU submission to epic-kitchens action anticipation challenge 2021. *CoRR*, abs/2107.13259, 2021. 2, 7

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[22] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022. 2, 3, 5, 6, 7

[23] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 5

[24] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2

[25] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8106–8116, 2021. 2

[26] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9925–9934, 2019. 2

[27] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020. 2

[28] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 2

[29] Shuang Li, Yilun Du, Antonio Torralba, Josef Sivic, and Bryan Russell. Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1845–1855, 2021. 2, 4

[30] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 2

[31] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 2

[32] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020. 1, 2, 7

[33] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022. 2

[34] Siyuan Brandon Loh, Debaditya Roy, and Basura Fernando. Long-term action forecasting using multi-headed attention-based variational recurrent neural networks. In *IEEE International Conference on Computer Vision and Pattern Recognition CVPR 2022 (Workshop)*, 2022. 2

[35] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023. 2

[36] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020. 2

[37] Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. Leveraging the present to anticipate the future in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[38] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021. 2, 6

[39] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. In *Advances in Neural Information Processing Systems*. 2

[40] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. Rethinking learning approaches for long-term action anticipation. In *European Conference on Computer Vision*, pages 558–576. Springer, 2022. 2

[41] Yan Bin Ng and Basura Fernando. Forecasting future action sequences with attention: a new approach to weakly supervised action forecasting. *IEEE Transactions on Image Processing 2020*, 2020. 2

[42] Yangjun Ou, Li Mi, and Zhenzhong Chen. Object-relation reasoning graph for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20133–20142, 2022. 2, 3, 5

[43] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021. 1, 2, 3, 4, 5, 6, 7

[44] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[45] Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3644, 2023. 2

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[47] Cristian Rodriguez, Basura Fernando, and Hongdong Li. Action anticipation by predicting future dynamic images. In *European Conference on Computer Vision*, pages 89–105. Springer, 2018. 2

[48] Debaditya Roy and Basura Fernando. Action anticipation using pairwise human-object interactions and transformers. *IEEE Transactions on Image Processing*, 30:8116–8129, 2021. 1, 2

[49] Debaditya Roy and Basura Fernando. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2745–2753, January 2022. 2

[50] Debaditya Roy and Basura Fernando. Predicting the next action by modeling the abstract goal. *arXiv preprint arXiv:2209.05044*, 2022. 2, 7

[51] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020. 2, 7

[52] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 2

[53] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13688–13697, 2021. 2

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[55] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018. 2

[56] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2, 7

[57] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2021. 2

[58] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. Multimodal temporal convolutional network for anticipating actions in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2249–2258, 2021. 2

[59] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021. 2

[60] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Is an object-centric video representation beneficial for transfer? *arXiv preprint arXiv:2207.10075*, 2022. 2

[61] Tianyu Zhang, Weiqing Min, Jiahao Yang, Tao Liu, Shuqiang Jiang, and Yong Rui. What if we could not see? counterfactual analysis for egocentric action anticipation. In *IJCAI*, 2021. 2

[62] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023. 2, 7