# Memory-and-Anticipation Transformer for Online Action Understanding

Jiahao Wang[1*] Guo Chen[1,2*] Yifei Huang[2] Limin Wang[1,2] Tong Lu[1✉]
[1] State Key Laboratory for Novel Software Technology, Nanjing University
[2] Shanghai AI Laboratory

## Abstract

*Most existing forecasting systems are memory-based methods, which attempt to mimic human forecasting ability by employing various memory mechanisms and have progressed in temporal modeling for memory dependency. Nevertheless, an obvious weakness of this paradigm is that it can only model limited historical dependence and can not transcend the past. In this paper, we rethink the temporal dependence of event evolution and propose a novel memory-anticipation-based paradigm to model an entire temporal structure, including the past, present, and future. Based on this idea, we present Memory-and-Anticipation Transformer (MAT), a memory-anticipation-based approach, to address the online action detection and anticipation tasks. In addition, owing to the inherent superiority of MAT, it can process online action detection and anticipation tasks in a unified manner. The proposed MAT model is tested on four challenging benchmarks TVSeries, THUMOS'14, HDD, and EPIC-Kitchens-100, for online action detection and anticipation tasks, and it significantly outperforms all existing methods. Code is available at* `https://github.com/Echo0125/Memory-and-Anticipation-Transformer`.

## 1. Introduction

Online anticipation [32] or detection [10] in computer vision attempt to perceive future or present states from a historical perspective. They are the crucial factors of AI systems that engage with complex real environments and interact with other agents, *e.g.* wearable devices [46], human-robot interaction systems [33], and autonomous vehicles [62].

Human beings frequently imagine future events based on past experiences. Similarly, anticipating future events inherently requires modeling past actions or the progression of events to predict what will happen next. To bridge the gap with human forecasting ability, current systems strive to mimic this cognitive ability by employing different memory mechanisms [58, 26, 7, 67, 44, 17]. Previous research [14, 44] demonstrated that modeling long temporal context is crucial for accurate anticipation. LSTR [58] further decomposes the memory encoder into long and short-term stages for online action detection and anticipation, allowing for the extraction of more representative memory features. Similarly, research efforts such as [67, 7, 27] have also illustrated improvements based on this principle.

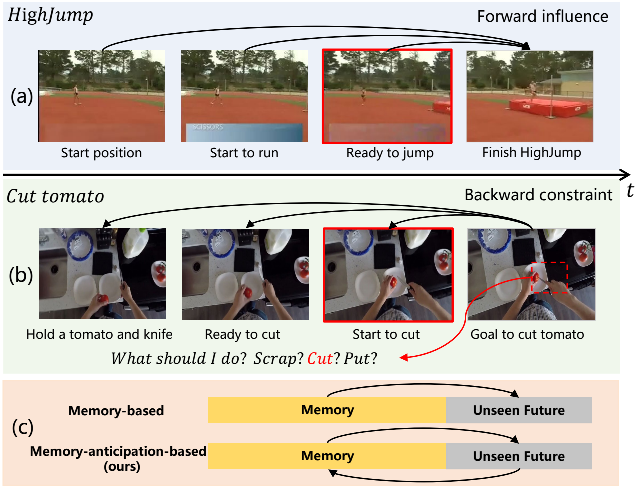

**Figure 1: Examples for the forward influence of memory and backward constraint of anticipation.** (a) An athlete prepares to start, starts to run, and then jumps to form the result of completing a high jump; (b) The goals of a camera wearer for cutting a tomato dictate the sequence of actions he will perform: hold the knife, get ready to cut, etc.; (c) Compared with the memory-based method, the memory-anticipation-based method is able to establish circular dependencies between memory and future.

However, despite various **memory-based** approaches, memory is not the only driver of present or future action. The synchronization between shot transitions and the evolution of actor behavior appears to cohere with the beliefs and wills of the event weavers, be they actors or camera-wearers. In (a) and (b) of Fig 1, we use the examples of high jumping and cooking to show the forward influence of memory and backward constraint of anticipation.

[*] Equal contribution, ✉ Corresponding author (lutong@nju.edu.cn)

Upon closer examination of both examples, it has been observed that future-oriented thoughts may impact action and memory, which seems modulated by the encoding of new information [43, 25]. Additionally, anticipation may change as ongoing behavior progresses and memory is updated [41, 20]. This indicates that there exists a circular interdependence between anticipation and memory, constraining the evolution of behavior or events.

Based on the preceding analysis, we reevaluate the temporal dependencies inherent in **memory-based** methods. In their preoccupation with the impact of memory on anticipation, these memory-based methods tend to overlook the inverse direction of impact, *i.e*., anticipation on memory. Consequently, historical representations are not adequately corrected, thereby risking impeding any attempts to transcend the past. Against this backdrop, we assert that a comprehensive temporal structure seamlessly integrates memory and anticipation is indispensable. In other words, a **memory-anticipation-based** method building circular dependencies between memory and anticipation, as shown in Fig 1(c), holds tremendous promise for enhancing cognitive inference capabilities and advanced understanding of AI systems about the present and future.

To this end, we propose **M**emory-and-**A**nticipation **T**ransformer **(MAT)**, a novel **memory-anticipation-based** approach that fully models the complete temporal context, including history, present, and future. A *Progressive Memory Encoder* is designed to provide a more precise history summary by compressing long- and short-term memory in a segment-based fashion. Meanwhile, we propose our key idea of modeling circular dependencies between memory and future, implemented as *Memory-Anticipation Circular Decoder*. It first *learns latent future features* in a supervised manner, then updates iteratively the enhanced short-term memory and the latent future features by performing *Conditional circular Interaction* between them. Among them, multiple interaction processes capture the circular dependency and supervise the output to maintain stable features with real semantics.

Remarkably, owing to the inherent superiority of our model design, we are able to adapt both tasks, *i.e*., online action detection and anticipation, in a unified manner, spanning the training and online inference stages. For any given dataset, the MAT model obviates the need for separate training or testing for each task. Rather, a single training process is enough, and during inference, the corresponding token for each task can be extracted effortlessly.

In summary, our contributions are 1). We rethink the temporal dependence of event evolution and propose a memory-anticipation-based paradigm for the circular interaction of memory and anticipation, introducing the concept of memory and future circular dependence. 2). We propose a unified architecture Memory-Anticipation Transformer (MAT) that simultaneously processes online action detection and anticipation, showing effective performance. 3). MAT significantly outperforms all existing methods on four challenging benchmarks for online action detection and anticipation tasks, *i.e*., TVSeries [10], THUMOS'14 [30], HDD [42] and EPIC-Kitchens-100 [9].

## 2. Related Work

**Online Action Detection and Anticipation.** Online action detection [10] aims to predict the action as it happens without accessing the future. Previous methods [57, 12, 16, 61, 54, 58, 7, 67, 4] for online action detection include recurrent networks, reinforcement learning, and, more recently, transformers. [15] uses a reinforced network to encourage making great decisions as early as possible. [57] adopts RNN to generate future predictions to improve online action detection. [58] proposes a transformer to scale the history spanning a longer duration. The target of action anticipation [9, 14, 17, 56] is predicting what actions will occur after a certain time. RULSTM [14] explores action anticipation and proposes an LSTM-based network to solve this problem. AVT [17] proposed an end-to-end attention-based model for anticipative video modeling and learns feature representations by self-supervised methods.

Online action detection and anticipation share two key characteristics: the inability to perceive future representations, with information derived solely from historical context, and the goal of predicting actions after a certain time. Previous research has typically studied the model for each task in isolation [61, 7] or trained and tested them separately [58, 67], ignoring the potential benefits of leveraging their shared characteristics. We propose a novel unified framework, named MAT, to address this limitation, which allows us to train or infer both tasks simultaneously.

**Transformer for Video Understanding.** Transformers have proven successful in natural language processing [49], and researchers have shown increasing interest in its application to vision tasks [11, 48, 40, 53, 18, 37]. Several methods have been proposed to use transformers for temporal modeling in video-related tasks [1, 2, 6, 28, 35, 36, 39, 52, 55, 65]. Timesformer [3] proposes divided attention to capture spatial and temporal information for action recognition separately. MViT [13] builds a feature hierarchy by progressively expanding channel capacity and reducing video spatial-temporal resolution. Actionformer [63] adopts the transformer architecture with local attention for temporal action localization. TubeDETR [60] tackles video grounding with an encoder-decoder transformer-based architecture to efficiently encode spatial and multi-modal interactions. Our MAT model is also built upon the transformer, utilizing its encoder and decoder architectures. We meticulously integrate the transformer blocks to implement our memory encoder and memory-anticipation circular decoder.

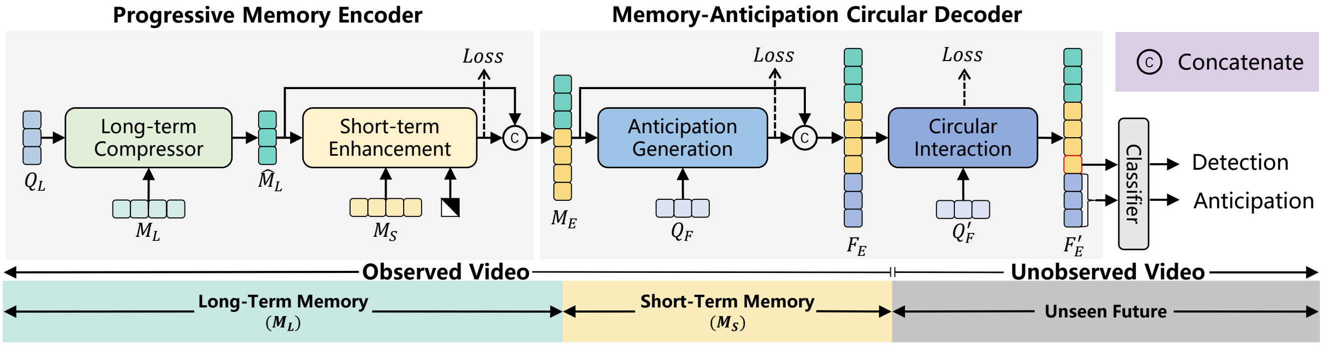

**Figure 2: MAT architecture.** We first divide the temporal sequence into long-term and short-term memories. In *Progessive Memory Encoder*, long-term memory queries $\mathbf{Q}_L$ progressively map the long-term to an abstract representation and be fed to the transformer decoder block to enhance the short-term memory. Then the *Memory-Anticipation Circular Decoder* utilizes learnable queries $\mathbf{Q}_F$ and $\mathbf{Q}'_F$ to perceive the future context and circularly updates the historical and future representation. Finally, a weight-shared classifier is adopted to output the classification scores of short-term and future for online action detection and anticipation.

## 3. Problem Setup

Online action anticipation and detection share a similar goal of predicting actions in untrimmed video sequences. In the former, we aim to forecast actions after a time gap $\tau$, as defined in [9], while in the latter, we predict the action under $\tau = 0$. Both tasks require classifying actions without access to future information during inference. We represent the input video as $\mathbf{V} = \{v_t\}_{t=-T+1}^{0}$, and our objective is to predict the action category $\hat{y}_\tau \in \{0, 1, 2, \ldots, C\}$ occurs after a specified time gap $\tau$ in the future. Here, $C$ represents the total number of action categories, and label 0 denotes the background category.

## 4. Memory-and-Anticipation Transformer

We now describe Memory-and-Anticipation Transformer (MAT) model architecture, as illustrated in Fig 2. It processes the memory cached in runtime to predict current or future actions. We adopt the *encoder-decoder* architecture to implement the model. The *Progressive Memory Encoder* is proposed to improve the quality of compressing long-term memory and enhancing short-term memory, which has a fundamental impact on anticipation and detection. MAT also employs our presented key component *Memory-Anticipation Circular Decoder* that generates the latent anticipation feature and conducts interaction between memory and anticipation in an iterative loop. We now introduce each model component in detail, followed by the training, online inference, and implementation details.

### 4.1. Progressive Memory Encoder

We use video feature $\mathbf{F} = \{f_t\}_{t=-T+1}^{0} \in \mathbb{R}^{T \times D}$ generated by a pretrained feature extractor as the input to our model, where $D$ and $T$ are the dimension and length of the feature sequence, respectively. To better handle the long memory, following [58], we divide the feature sequence into two consecutive memories. The first is the short-term memory that stores only a handful of recently occurred frames. It stores the feature vectors as $\mathbf{M}_S = \{f_t\}_{t=-m_S+1}^{0}$, where $m_S$ denotes the length. The other, named long-term memory, contains the feature that is far away from the current time, which is defined as $\mathbf{M}_L = \{f_t\}_{t=-T+1}^{-m_S}$. We set $m_L$ to be the length of the long-term memory, which is much longer than the short-term memory.

In practice, *Progressive Memory Encoder* progressively encodes long- and short-term memories. It first compresses the long-term memory to an abstract representation and injects them into the short-term memory. Differing with previous works [58, 67] that separately tackle the long- and short-term memories by one long-term encoder and one short-term decoder, we integrate these two works into a single encoder architecture. It is due to our encoded short-term memory being served for decoders and not as the final output. In our experiments, we observe that the design of the memory encoder is crucial for model performance. Therefore, different with [58, 67], we introduce a novel long-term memory compression method to improve the quality of memory encoding.

**Segment-based Long-term Memory Compression.** LSTR [58] adopts a two-stage compression technique to project the long-term memory to a fixed-length latent embedding. It implements a function similar to the bottleneck layer [22]. Long-term memory, however, represents a long token sequence that often carries much noise. It is problematic to generate attention weights while cross-attention queries too many tokens with mixed noise. We propose *Segment-based Long-term Memory Compression* to alleviate the issue, as shown in Fig 3. Concretely, we first divide $\mathbf{M}_L$ into $N_s$ non-overlapping memory segments $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^{N_s}$ of length $\frac{m_L}{N_s}$. We then employ $N_L$ learnable tokens $\mathbf{Q}_L \in \mathbb{R}^{N_L \times D}$ as the long-term memory queries and a weight-shared transformer decoder block to query each segment. $\mathbf{S}$ will be transformed to $N_s$ segment-level abstract feature $\mathbf{F} = \{\mathbf{f}_i\}_{i=1}^{N_s}$, where $\mathbf{f}_i \in \mathbb{R}^{N_L \times D}$. We average pool each $\mathbf{f}_i$ to one vector $\mathbb{R}^D$ and concatenate them to form the long-term compressed segmented memory $\mathbf{M}_L^s \in \mathbb{R}^{N_s \times D}$. At last, we feed them into two transformer encoder blocks to obtain the final compressed long-term memory $\widehat{\mathbf{M}}_L \in \mathbb{R}^{N_s \times D}$.

**Short-term Memory Enhancement.** As short-term memory $\mathbf{M}_S = \{f_t\}_{t=-m_S+1}^{0}$ contains crucial trends for accurately predicting the current or upcoming activities, we leverage short-term memory as the query to retrieve relevant context from the compressed long-term memory. As illustrated in Fig 2, we employ a transformer causal decoder block to aggregate compressed long-term memory $\widehat{\mathbf{M}}_L$ into short-term memory. The enhanced short-term memory $\widehat{\mathbf{M}}_S$ is then used to identify actions at each timestep, constituting a sequence of previously occurring actions and predicting future states and actions. We use frame-level action labels to supervise the dense classification task, encouraging the model to produce a clear sequence of historical behavior.

## 4.2. Memory-Anticipation Circular Decoder

To implement our key idea that considers the co-reaction between anticipation and memory, we carefully design *Memory-Anticipation Circular Decoder* that circularly performs interaction between them. It first generates the latent anticipation that is a prerequisite for memory-anticipation interaction. Then, *Conditional Circular Interaction* will generate new memories and anticipation conditioned on old ones iteratively. Finally, a weight-shared classification head will supervise the generated memories and anticipation.

**Latent Anticipation Generation.** Initially, we do not have an anticipation state. Therefore, we must first generate a latent anticipation feature, *i.e.*, a future representation containing abstract information derived from known memory. Given the compressed long-term memory $\widehat{\mathbf{M}}_L$ and short-term memory $\widehat{\mathbf{M}}_S$, we first concatenate them to form the entire memory $\mathbf{M}_E = [\widehat{\mathbf{M}}_L, \widehat{\mathbf{M}}_S]$, where $[\cdot, \cdot]$ is the concatenating operation along the temporal dimension. Then, we use $N_F$ future queries $\mathbf{Q}_F \in \mathbb{R}^{N_F \times D}$ and a transformer decoder block to query $\mathbf{M}_E$. The updated future queries are represented as the latent anticipation features $\mathbf{F}_A \in \mathbb{R}^{N_F \times D}$ at the current state. We use future information, *e.g.*, future video features and future action labels, to supervise $\mathbf{F}_A$. In practice, we use information from the next $T_F$ seconds. Since $\mathbf{F}_A$ is only used to describe the latent future states, too large future embeddings may overfit the anticipation representation. We set $N_F$ to be smaller than the length of the used future information sequence. Then we use bilinear interpolation to upsample it to align for point-to-point supervision. Note that although we use future information here, this does not lead to an information leak. The model will generate preliminary anticipation in the inference phase and not use future labels or features.

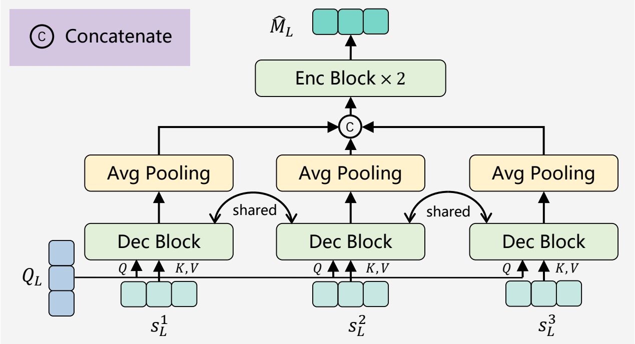

**Figure 3: Segment-based Long-term Memory Compression.** For example, given $m_L = 9$, $N_s = 3$, the process is shown above. We can finally obtain compressed features $\widehat{M}_L$ of length 3.

**Conditional Circular Interaction.** So far, we have generated long-term memory $\widehat{\mathbf{M}}_L$, short-term memory $\widehat{\mathbf{M}}_S$ and latent anticipation features $\mathbf{F}_A$. The memories, however, are limited to the past and can not go beyond history. We now describe how *Conditional Circular Interaction* works. For brevity, we detail how memory and anticipation interact with each other one time.

As shown in Fig 4, we first concatenate $\widehat{\mathbf{M}}_L$, $\widehat{\mathbf{M}}_S$ and $\mathbf{F}_A$ to form the entire feature $\mathbf{F}_E = [\widehat{\mathbf{M}}_L, \widehat{\mathbf{M}}_S, \mathbf{F}_A]$. Then, $\mathbf{F}_E$ and $\widehat{\mathbf{M}}_S$ are feed into a transformer decoder block:

$$\widehat{\mathbf{M}}'_S = \text{CrossAttn}(\widehat{\mathbf{M}}_S, \mathbf{F}_E, \mathbf{F}_E), \tag{1}$$

where $\text{CrossAttn}(Q, K, V)$ is the cross-attention layer in the transformer decoder block. The updated short-term memory is represented as $\widehat{\mathbf{M}}'_S$. Since $\mathbf{F}_E$ contains memories and anticipation, $\widehat{\mathbf{M}}_S$ as a query condition dynamically extracts new semantics mixed with memories and anticipation from three representations. Next, we re-concatenate $\widehat{\mathbf{M}}_L$, $\widehat{\mathbf{M}}'_S$ and $\mathbf{F}_A$ to generate the new entire feature $\mathbf{F}'_E = [\widehat{\mathbf{M}}_L, \widehat{\mathbf{M}}'_S, \mathbf{F}_A]$. Another transformer decoder block is used to query $\mathbf{F}'_E$, with $\mathbf{F}_A$ as the query condition:

$$\mathbf{F}'_A = \text{CrossAttn}(\mathbf{F}_A, \mathbf{F}'_E, \mathbf{F}'_E), \tag{2}$$

where $\mathbf{F}'_A$ is the resulting anticipation features. After that, $\mathbf{F}'_A$ is injected with the higher-order inferred semantics brought about by the interaction between memory and anticipation.

In the interaction process, the generated $\widehat{\mathbf{M}}'_S$ is both supervised like the above initial anticipation features. It should be noted that different from generating latent initial anticipation through fewer future queries, we find that using a new group of future queries $\mathbf{Q}'_F \in \mathbb{R}^{N'_F \times D}$ strictly aligned future information sequences of $T_F$ seconds in the first interaction process to replace the $\mathbf{F}_A$ with $\mathbf{Q}'_F$, bring

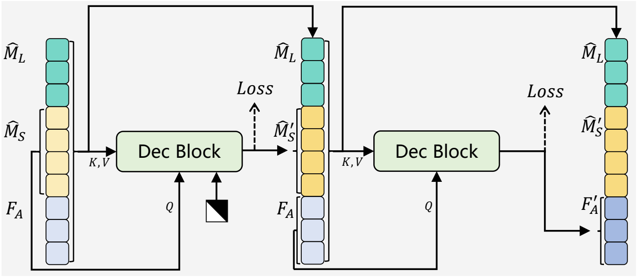

**Figure 4: Conditional circular interaction** dynamically updates and re-constructs the context between memories and anticipation by aggregating different temporal information.

performance improvement. Thus, we point-to-point supervise the anticipation $\mathbf{F}'_A$ generated by each interaction process. Meanwhile, this design is more natural to output the corresponding token according to the anticipation gap time. We circularly perform the interaction process $N_t$ times until saturation. The yielded $\widehat{\mathbf{M}}'_S$ and $\mathbf{F}'_A$ in one interaction process become the new $\widehat{\mathbf{M}}_S$ and $\mathbf{F}_A$ and are passed to the next interaction process.

### 4.3. Training

Our MAT model relies on action steps of short-term memory to predict the current or future action. These steps are utilized as auxiliary information to benefit action detection and anticipation. It is worth noting that since the MAT model regards online action detection and anticipation as the same tasks with different settings, we can define the loss function in a unified fashion.

For the $\widehat{\mathbf{M}}_S$ generated by the encoder, we feed it to a classifier to generate the action probabilities $\widehat{\mathbf{Y}}_S = \{\widehat{\mathbf{y}}^i_S\}^{m_S}_{i=1}$. We then supervise the whole short-term memory using a cross-entropy loss with the labeled action, $\{c^i_S\}^{m_S}_{i=1}$. Learned future features $\mathbf{F}_A$ are fed into the same classifier to generate the probabilities $\widehat{\mathbf{Y}}_F = \{\widehat{\mathbf{y}}^i_F\}^{N_F}_{i=1}$ and we also adopt a cross-entropy loss with the target, $\{c^i_F\}^{m_S}_{i=1}$:

$$\mathcal{L}^0_S = -\sum_{i=1}^{m_S} \log \widehat{\mathbf{y}}^i_S[c^i_S], \mathcal{L}^0_F = -\sum_{i=1}^{N_F} \log \widehat{\mathbf{y}}^i_F[c^i_F], \tag{3}$$

where $\mathcal{L}^0_S$ and $\mathcal{L}^0_F$ are the loss function of encoded short-term memory and initial anticipation features, respectively. As the conditional circular interaction is conducted $N_t$ times, we use $\mathcal{L}^i_S$ and $\mathcal{L}^i_F$ to represent the loss of memory and anticipation in each interaction process, where $1 \leq i \leq N_t$. Their calculation is similar to $\mathcal{L}^0_S$ and $\mathcal{L}^0_F$. The final training loss is formulated by:

$$\mathcal{L} = \sum_{i=0}^{N_t} \lambda^i_s \cdot \mathcal{L}^i_S + \lambda_f \sum_{i=0}^{N_t} \mathcal{L}^i_F, \tag{4}$$

where $\lambda^i_s$ and $\lambda_f$ are the balance coefficients.

### 4.4. Online Inference

The existing works [57, 61, 58, 67, 17] test each task independently. Thanks to the design of the unified framework, MAT simultaneously infers online action detection and anticipation tasks on one streaming video. It uses the output $\widehat{\mathbf{M}}_S$ and $\mathbf{F}_A$ produced by the last interaction as results of online action detection and anticipation, respectively. We take out the last token of $\widehat{\mathbf{M}}_S$ for online action detection. For action anticipation, according to gap time $\tau$, we take out the corresponding token from $\mathbf{F}_A$ for forecasting.

## 5. Experiments

### 5.1. Dataset and Metrics

**Dataset.** We test on three online action detection datasets. *THUMOS'14* [30] consists of over 20 hours of sports video annotated with 20 actions. *TVSeries* [10] includes 27 episodes of 6 popular TV series, about 150 minutes each and 16 hours total. *HDD* [42] is a large-scale human driving video dataset comprising 104 hours of untrimmed videos and 11 action categories. Furthermore, we evaluate on *EpicKitchens-100* [9] that contains 100 hours of egocentric videos with at least 90K action segments and whose narrations are mapped to 97 verb classes and 300 noun classes.

**Evaluation Metrics.** For online action detection, following previous works [10, 34, 54, 58, 61], we use per-frame *mean Average Precision (mAP)* on THUMOS'14 and HDD, and per-frame *mean calibrated Average Precision (mcAP)* [10] on TVSeries. For action anticipation, we employ *Top-5 Verb/Noun/Action Recall* to measure the performance with an anticipation period $\tau = 1s$ [9].

### 5.2. Implementation details

Following prior works [57, 58, 67], we conduct our experiments on pre-extracted features. For TVSeries and THUMOS'14, we first resample the videos at 24 FPS and then extract the frames at 4 FPS for training and validation. We adopt a two-stream network [50] to extract feature. We use off-the-shell checkpoint released by mmaction2 [8] that pretrained on ActivityNet v1.3 [23] and Kinetics [31] to extract frame-level RGB and optical features. On EK100, following [14], we resample the videos at 30 FPS and then fine-tune the two-stream TSN on the classification task.

**Memory Dropping.** Despite the promising capabilities of the dense attention mechanism, capturing long-range dependencies remains a challenge. However, relying on the dense attention mechanism may lead to sub-optimization of the network, resulting in a locally optimal solution. We seek to explore memory-dropping strategies for implementing sparse attention during training. We experiment with various approaches, including dropout [24], top-k selec-

| $N_s$ | mAP |
|---|---|
| [58] | 69.6 |
| 2 | 69.8 |
| 4 | 70.2 |
| 8 | **70.4** |
| 12 | 70.1 |

**(a) Segment number.**

| renewal | mAP |
|---|---|
| 0 | 70.0 |
| 1 | **70.4** |
| 2 | 69.7 |

**(b) Future queries renewing.**

| deep | supervision | mAP |
|---|---|---|
| - | cls | 69.9 |
| ✓ | feat | 69.6 |
| ✓ | cls | **70.4** |
| ✓ | feat + cls | 70.2 |

**(c) Future supervision.**

| short | future | mAP |
|---|---|---|
| unshared | | 69.6 |
| separately shared | | 70.1 |
| full shared | | **70.4** |

**(d) Shared classifier.**

| long | short | V | N | A |
|---|---|---|---|---|
| - | - | 32.1 | 36.1 | 18.1 |
| MC | - | 32.9 | 36.9 | 18.4 |
| MC | MC | 31.0 | 37.4 | 18.5 |
| MC | MC+ | **35.0** | **38.8** | **19.5** |
| MC+ | MC+ | 31.6 | 36.8 | 18.4 |

**(e) MixClip+.**

**Table 1: Ablation Experiments.** We conduct detailed ablation on (a): Segment number, (b): Future queries renewing, (c): Future supervision, (d): Shared classifier, and (e): MixClip+. The gray rows denote default choices.

| | long | short | future | interaction | $N_t$ | detection | anticipation |
|---|---|---|---|---|---|---|---|
| (a) | | | | - | - | $67.6_{+0.0}$ | $53.7_{+0.0}$ |
| (b) | | ✓ | | CA | 1 | $68.3_{+0.7}$ | $54.7_{+1.0}$ |
| (c) | ✓ | ✓ | | CA | 1 | $68.9_{+1.3}$ | $55.3_{+1.6}$ |
| (d) | ✓ | | ✓ | CA | 1 | $68.6_{+1.0}$ | $55.8_{+2.1}$ |
| (e) | | ✓ | ✓ | CA | 1 | $69.2_{+1.6}$ | $56.1_{+2.4}$ |
| (f) | ✓ | ✓ | ✓ | Avg&Cat | 1 | $68.3_{+0.7}$ | $55.3_{+1.6}$ |
| (g) | ✓ | ✓ | ✓ | CA | 1 | $69.9_{+2.3}$ | $56.6_{+2.9}$ |
| (h) | ✓ | ✓ | ✓ | CA | 2 | $\mathbf{70.4}_{+2.9}$ | $\mathbf{57.3}_{+3.6}$ |
| (i) | ✓ | ✓ | ✓ | CA | 3 | $70.2_{+2.6}$ | $57.1_{+3.4}$ |

**Table 2: Ablation study** on the information stream, interaction method, and times for Conditional Circular Interaction. The shallow gray row denotes the baseline setting without any interaction.

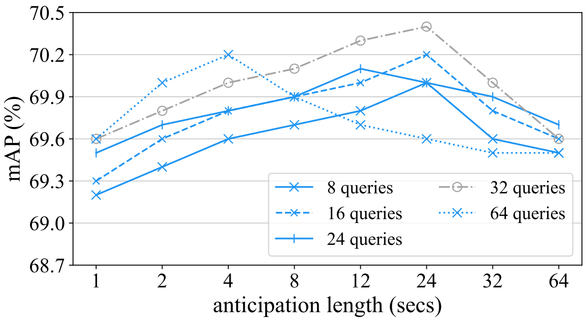

**Figure 5: Generating latent anticipation** with different anticipation length $T_F$ and numbers of learnable tokens $N_F$.

tion [51], and token dropping [21]. We compare these strategies, recorded in the supplementary material, and find that the top-k selection yields the best performance.

**MixClip+.** TesTra [67] proposed an augmentation technique called MixClip to solve over-fitting caused by large long-term memory. Unlike long-term memory which presents an abstract concept, short-term memory provides a continuous motion trend, often playing a decisive role in anticipation. It leads to the fact that MixClip damages the continuity of motion, thereby, does not apply to short-term memory. We propose MixClip+ to solve the issue.

In addition to mixing augmentation for long-term memory, we mix each feature token of short-term memory with a random clip from different videos. To maintain the continuity of short-term memory, we adopt soft fusion with hyperparameter $\alpha = 0.25$ to mix features and the corresponding labels, similar to Mixup [64]. Due to space limitations, we will describe this part in the supplementary material.

## 5.3. Ablation Study

Now we analyze the MAT model, using the two-stream features pretrained on ActivityNet v1.3 and THUMOS'14 test set as the test bed. Furthermore, we also study the contribution of the proposed MixClip+ on EPIC-Kitchens-100.

**Segment Number.** We evaluate the effect of segment number $N_s$ for long-term memory compression in Table 1a. We find even a small $N_s$ can improve performance, compared with two-stage compression [58]. However, too many segments result in too little information within each segment to effectively capture temporal dependency. We use $N_s = 8$ as the default in the following ablation experiments.

**Query Number and Anticipation Length.** Fig 5 compares generating latent anticipation using $N_F$ learnable tokens with different anticipation lengths $T_F$. The experimental results indicate that, for $T_F \leq 4$, a larger $N_F$ yields superior results owing to strongly correlated semantics of short-term memory and short-term anticipation thus, the model can benefit from more tokens. We observe that $N_F = 8, 16, 32$ exhibits a similar trend in the accuracy curve, with the model achieving optimal performance with $N_F = 32$. This suggests that long-term anticipation is more adaptable for sparse tokens. The observations imply that anticipation is also suitable for long- and short-term modeling.

**Future Queries Renewal.** As Section 4.2 outlines, renewing future queries before the initial interaction improves performance. In Table 1b, we compare different renewing times ("0" denotes no renewal and utilizes $F_A$). A plausible explanation is that, in the first interaction process, the interaction from anticipation to memory adjusts the old memory, causing its domain to shift largely, which leads to insufficient interaction from the new memory to the initial anticipation feature.

**Future Supervision.** In addition, we carry out diverse experiments to determine the optimal choice of future su-

| Method | Modality | Init | Overall |  |  | Unseen |  |  | Tail |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action |
| RULSTM [14] | RGB | IN-1K | 27.5 | 29.0 | 13.3 | 29.8 | 23.8 | 13.1 | 19.9 | 21.4 | 10.6 |
| AVT [17] |  | IN-1K | 27.2 | 30.7 | 13.6 | - | - | - | - | - | - |
| AVT [17] |  | IN-21K | 30.2 | 31.7 | 14.9 | - | - | - | - | - | - |
| DCR [59] |  | IN-1K | 31.0 | 31.1 | 14.6 | - | - | - | - | - | - |
| TeSTra [67] |  | IN-1K | 26.8 | 36.2 | 17.0 | 27.1 | 30.1 | **13.3** | 19.3 | 28.6 | 13.7 |
| MAT (Ours) |  | IN-1K | **32.7** | **39.7** | **18.8** | **31.7** | **32.1** | 12.7 | **25.7** | **32.4** | **16.0** |
| RULSTM [14] | RGB + OF + OBJ | IN-1K | 27.8 | 30.8 | 14.0 | 28.8 | 27.2 | 14.2 | 19.8 | 22.0 | 11.1 |
| TempAgg [44] |  | IN-1K | 23.2 | 31.4 | 14.7 | 28.0 | 26.2 | **14.5** | 14.5 | 22.5 | 14.8 |
| AVT+ [17] |  | IN-1K | 25.5 | 31.8 | 14.8 | 25.5 | 23.6 | 11.5 | 18.5 | 25.8 | 12.6 |
| AVT+ [17] |  | IN-21K | 28.2 | 32.0 | 15.9 | 29.5 | 26.0 | 12.8 | 23.2 | 29.2 | 14.1 |
| TeSTra [67] | RGB + OF | IN-1K | 30.8 | 35.8 | 17.6 | 29.6 | 26.0 | 12.8 | 23.2 | 29.2 | 14.2 |
| MAT (Ours) |  | IN-1K | **35.0** | **38.8** | **19.5** | **32.5** | **30.3** | 13.8 | **28.7** | **33.1** | **16.9** |

**Table 3: Comparison to prior work on EPIC-Kitchens-100 Action Anticipation [9].** Accuracy measured by class-mean recall@5(%) following the standard protocol.

| Modality | Encoder | Init | FT | Overall | Unseen | Tail |
|---|---|---|---|---|---|---|
| RGB | TSN | IN-1K | ✓ | 18.8 | 12.7 | 16.0 |
| RGB + OF | TSN | IN-1K | ✓ | 19.5 | 13.8 | 16.9 |
| RGB | ViT-L | K400 |  | 19.1 | 16.2 | 16.1 |
| RGB (V) | ViT-L | Ego4D |  | 21.9 | 19.4 | 18.5 |
| RGB (N) | ViT-L | Ego4D |  | 24.2 | 21.4 | 20.9 |
| RGB (V + N) | ViT-L | Ego4D |  | 24.6 | 21.4 | 21.2 |

**Table 4: Exploring different visual encoders** for EK100 action anticipation. V or N denotes the encoder is pretrained on verb or noun subset [5]. "FT" is that the encoder is fine-tuned on the classification task of EK100.

| Method | THUMOS'14 |  | TVSeries |  |
|---|---|---|---|---|
|  | Kinetics | ANet | Kinetics | ANet |
| RED [15] | - | 37.5 | 75.1 | - |
| TRN [57] | - | 38.9 | 75.7 | - |
| OadTR [54] | 53.5 | 45.9 | 77.8 | 79.1 |
| LSTR [58] | 52.6 | 50.1 | 80.8 | - |
| GateHUB [7] | - | 54.2 | 82.0 | - |
| TesTra [67] | 56.8 | 55.3 | - | - |
| MAT (ours) | **58.2** | **57.3** | **82.6** | **81.5** |

**Table 5: Action anticipation result** on THUMOS'14 and TVSeries, mAP is reported for THUMOS'14 and mcAP for TVSeries.

pervision, as illustrated in Table 1c. We investigate two methods, namely future feature supervision and future label supervision. The results suggest that, compared to feature supervision, label supervision yields superior accuracy by incorporating stronger semantics derived from manual annotation. Furthermore, we explore the significance of deep supervision, *i.e.*, multi-stage supervision. We observe that introducing multiple supervisions of real action labels improves performance.

**Shared Classifier.** In Table 1d, we exploit the classifier design in deep supervision for short-term memory and anticipation. We observe that sharing classifier parameters separately for them yields better performance. When we use a full-shared classifier for memory and anticipation, sharing all feature samples between them leads to the best performance. These results indicate that all yielded features in the interaction processes enrich the feature samples to be classified, reaching a function similar to the data augmentation.

**Conditional Circular Interaction.** Table 2 compares conditional circular interaction incorporating different interaction information streams, methods, and times. From (a) to (e), our experimental results demonstrate that the model's classification performance improves with the increasing richness of interaction information. Moreover, we observe that latent information in the future is equally or even more critical to classify the current action than information from the distant past, as evidenced by the performance comparison between (c) and (e). We also investigate the effect of different interaction mechanisms, comparing (f) and (g), and the results indicate that the cross-attention (CA) mechanism outperforms average pooling and concatenation similar to [54], achieving a 2.1% higher accuracy. Lastly, we evaluate the impact of different interaction times $N_t$, as illustrated by (g), (h), and (i). The experimental results reveal that the model's performance improves gradually with the increase of $N_t$. Notably, the model attains its highest performance when interacting with all information $N_t = 2$ times.

**MixClip+ for Anticipation.** Table 1e presents the augmentation efficacy of MixClip+ or MixClip on long- and short-term memory. Notably, the baseline experiences a significant drop in action recall to 18.1% without data augmentation. However, the result is still significantly better than Testra [67] with MixClip (18.1% *vs* 17.6%), as shown in Table 3. Furthermore, the results in row 3 reveal that MixClip

| Method | Augmentation | | Overall | | | Unseen | | | Tail | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Long | Short | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action |
| LSTR$^{\ddagger}$ [58] | - | - | 39.6 | 44.1 | 22.6 | 34.3 | 35.3 | 18.7 | 39.0 | 41.6 | 20.7 |
| TesTra$^{\ddagger}$ [67] | - | - | 40.0 | 44.8 | 23.2 | 34.6 | 36.0 | 19.0 | 39.4 | 42.1 | 20.9 |
| MAT (ours) | - | - | **41.8** | **46.1** | **24.9** | **36.5** | **37.9** | **20.1** | **40.8** | **43.2** | **22.8** |
| TesTra$^{\ddagger}$ [67] | MixClip | - | 39.7 | 45.6 | 25.1 | 36.3 | 37.2 | 19.2 | 39.0 | 42.4 | 22.2 |
| MAT (ours) | MixClip | - | 42.6 | 47.3 | 25.9 | 38.3 | 37.6 | 19.7 | 41.8 | 44.0 | 23.1 |
| MAT (ours) | MixClip | MixClip+ | **44.5** | **48.3** | **26.3** | **39.9** | **38.1** | **20.3** | **43.4** | **46.6** | **23.7** |

**Table 6: Online action detection result** on EPIC-Kitchens-100. Accuracy is measured by class-mean recall@5(%). $^{\ddagger}$ was reproduced by us because LSTR [58] and TesTra [67] did not report the result.

| Method | Arch | THUMOS'14 | | TVSeries | | HDD |
|---|---|---|---|---|---|---|
| | | Kinetics | ANet | Kinetics | ANet | Sensor |
| CDC [45] | CNN | - | 44.4 | - | - | - |
| RED [15] | RNN | - | 45.3 | - | 79.2 | 27.4 |
| TRN [57] | RNN | 62.1 | 47.2 | 86.2 | 83.7 | 29.2 |
| IDN [12] | RNN | 60.3 | 50.0 | 86.1 | 84.7 | - |
| LAP [34] | RNN | - | 53.3 | - | 85.3 | - |
| PKD [66] | CNN | 64.5 | - | 86.4 | - | - |
| WOAD [16] | RNN | 67.1 | - | - | - | - |
| OadTR [54] | Trans | 65.2 | 58.3 | 87.2 | 85.4 | 29.8 |
| Colar [61] | Trans | 66.9 | 59.4 | 88.1 | 86.0 | 30.6 |
| LSTR [58] | Trans | 69.5 | 65.3 | 89.1 | 88.1 | - |
| GateHUB [7] | Trans | 70.7 | 69.1 | 89.6 | 88.4 | 32.1 |
| TeSTra [67] | Trans | 71.2 | 68.2 | - | - | - |
| MAT (Ours) | Trans | **71.6** | **70.4** | **89.7** | **88.6** | **32.7** |

**Table 7: Online action detection performances** on THUMOS'14 [30], TVSeries [10], and HDD [42]. The mAP performance is reported for THUMOS'14 and HDD, while the mcAP is reported for TVSeries.

is not suited for short-term memory, as it cannot improve performance and may even degrade the results regarding the verb (32.1% *vs* 31.0%). However, the model achieves the best performance (19.5%) when performing MixClip and MixClip+ augmentation separately on long- and short-term memory. Moreover, we validate the augmentation of double-MixClip+ (in row 5) on long- and short-term memory and observe that the overall gain of the model is small, similar to double-MixClip (in row 3).

### 5.4. Comparison with State-of-the-Art Methods

**Action Anticipation.** We compare MAT with prior methods on EPIC-Kitchens-100 for action anticipation, as shown in Table 3. We divide the experimental results into two parts, one part of the input modality is RGB, and the other is multi-modal input, such as optical flow and object features. Using the ImageNet-1K pretrained features, MAT significantly outperforms RULSTM [14], AVT [17], and TeSTra [67] on terms of verb, noun, and action (at least 2.5%, 3.0%, and 1.8%, respectively). MAT with RGB and optical flow features achieves 1.8% higher action recall than TeSTra with the same input, which can better reflect the effectiveness of the modules we designed.

In Table 4, we conduct an extensive analysis of the impact of various visual encoders on the performance of action anticipation. In particular, we leverage the ViT architecture [11], which is pre-trained on the largest egocentric dataset Ego4D [19], made available by [5], owing to its superior performance in Ego4D Challenges. Additionally, to compare the impact of first-person and third-person pretraining, we utilize the weights pretrained on the K400 [31], provided by VideoMAE [47]. The results suggest that encoders pre-trained on first-person datasets perform better than on widely-used third-person datasets. Moreover, leveraging the noun subset brings a huge performance boost, comparing pretraining on the verb subset of the Ego4D dataset. We suppose the noun subset with more object categories (the verb and noun subset include 118 and 582 categories) encourages the encoder to search more complex visual semantics. It highlights the critical role of egocentric perspective upstream pretraining method [38, 68] and diverse semantic annotations [19, 9, 29], which could have far-reaching implications for egocentric research.

Following previous works [58, 7, 67], we further evaluate MAT on action anticipation tasks for THUMOS'14 and TVSeries. Unlike the previous work that required additional learnable tokens, we take the corresponding tokens from $\mathbf{F}_A$ for forecasting. Table 5 shows that using the ActivityNet pretrained features, MAT significantly outperforms the existing methods by 2.0% and 0.6% on THUMOS'14 and TVSeries, respectively. Moreover, the performance of MAT can also be further improved when pretraining on Kinetics.

**Online Action Detection.** We also compare MAT with other state-of-the-art online action detection methods [57, 58, 54, 61, 7] on THUMOS'14, TVSeries, and HDD. As illustrated in Table 7, our MAT achieves state-of-the-art performance and improves mAP by 2.2% on the THUMOS'14 dataset under the TSN-Anet feature input, and 0.4% under the TSN-Kinetics feature. For TVSeries, MAT is slightly better than GateHUB by 0.1% Furthermore, MAT achieves 0.6% better than the state-of-the-art GateHUB (32.1%) method on HDD.

To further verify the effect of MAT, in Table 6, we offer

| Method | #Param | FPS | | | | | mAP |
|---|---|---|---|---|---|---|---|
| | | Optical Flow | RGB Feat | Flow Feat | Model | Total | |
| TRN | 402.9 M | 8.1 | 70.5 | 14.6 | **123.3** | 8.1 | 47.2 |
| OadTR | 75.8 M | 8.1 | 70.5 | 14.6 | 110.0 | 8.1 | 58.3 |
| LSTR | 58.0 M | 8.1 | 70.5 | 14.6 | 91.6 | 8.1 | 65.3 |
| GateHUB | 45.2 M | 8.1 | 70.5 | 14.6 | 71.2 | 8.1 | 69.1 |
| MAT (Ours) | 94.6 M | 8.1 | 70.5 | 14.6 | 72.6 | 8.1 | **70.4** |

**Table 8: Efficiency comparison** between MAT and the previous work on parameter (M) and inference speed (FPS)

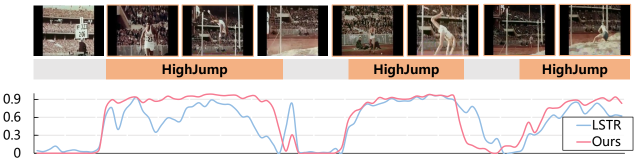

**Figure 6: Visualization of MAT's online prediction.** The curves indicate the predicted confidence of the ground-truth class (*HighJump*) with LSTR and our method.

the results for online action detection on the EK100 dataset. We divide the experimental results into two parts: one does not use augmentation, and the other uses MixClip or MixClip+. The results show that MAT surpasses the previous method by a large margin whether or not to use data augmentation. It is worth noting that when performing online action detection, we only need to take out the corresponding token in the short-term memory $\widehat{\mathbf{M}}_S$ for prediction. The superior performance of the model further highlights that MAT can simultaneously handle online action detection and anticipation tasks, thus integrating the two problems into a unified framework and performing excellently.

### 5.5. Efficiency Analysis

Table 8 compares MAT with other methods in terms of model parameters and running time, conducted on a single Tesla V100. It is worth noting that under the same features, the efficiency bottleneck of the system remains in optical flow calculation and feature extraction. Our approach achieves a stronger performance while effectively balancing the trade-off between parameters and computational.

### 5.6. Qualitative Analysis.

In Fig 6, we qualitatively analyzes the proposed MAT model. The confidence in the range $[0, 1]$ on the y-axis denotes the probability of predicting the current action,*e.g.*, *HighJump*. The figure highlights our MAT's successful suppression of background and high-confidence feedback of action segments compared with LSTR. The circular interaction of memory and anticipation provides reliable semantic information for the online prediction of the model, making the model offers a more sensitive and accurate torsion signal on the transition between background and foreground.

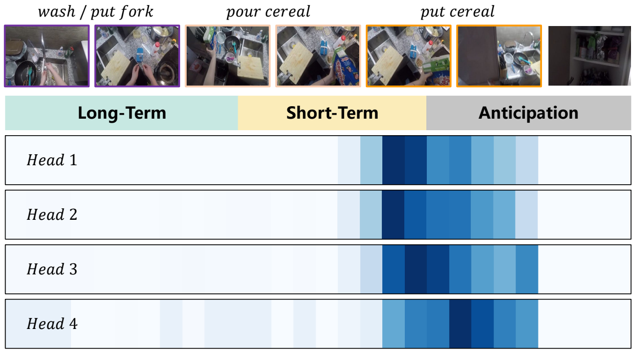

**Figure 7: Attention weight visualization of different attention heads.**

### 5.7. Attention Visualization

Fig 7 depicts the attention weights of the last token of the short-term memory. It can be seen that under the blessing of future perception, MAT's multi-head attention mechanism pays attention to memory and anticipation simultaneously, leading to the model not limiting in history and attaining better performance.

## 6. Conclusion and Future Work

We present Memory-and-Anticipation Transformer (MAT), a novel memory-anticipation-based paradigm for online action detection and anticipation, to overcome the weakness of most existing methods that can only complete modeling temporal dependency within a limited historical context. Through extensive experiments on four challenging benchmarks across two tasks, we show its applicability in predicting present or future actions, obtaining state-of-the-art results, and demonstrating the importance of circular interaction between memory and anticipation in the entire temporal structure. Although the temporal dimension is only considered in this work, we believe MAT would be a general paradigm for any AI system used to analyze video online and predict future events. In the near future, we will validate our model on more benchmarks and extend it to the long-term anticipation task. In the longer term, we will continue to unearth the intrinsic association of memory with anticipation and develop more effective forecasting models.

## Acknowledgements

This work was supported in part by National Key R&D Program of China (No. 2022ZD0160900), National Natural Science Foundation of China (Grant No. 61672273, 61832008, 62076119), and was also supported by the Fundamental Research Funds for the Central Universities XJ2023000701 and XJ2023004101.

## References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6816–6826, 2021. 2

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. 2

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 139, pages 813–824, 2021. 2

[4] Shuqiang Cao, Weixin Luo, Bairui Wang, Wei Zhang, and Lin Ma. A circular window-based cascade transformer for online action detection. *CoRR*, abs/2208.14209, 2022. 2

[5] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, Zhiyu Zhao, Junting Pan, Yifei Huang, Zun Wang, Jiashuo Yu, Yinan He, Hongjie Zhang, Tong Lu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *CoRR*, abs/2211.09529, 2022. 7, 8

[6] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. Videollm: Modeling video sequence with large language models. *CoRR*, abs/2305.13292, 2023. 2

[7] Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, and Mei Chen. Gatehub: Gated history unit with background suppression for online action detection. In *CVPR*, pages 19925–19934, 2022. 1, 2, 7, 8

[8] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 5

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2, 3, 5, 7, 8

[10] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *ECCV*, pages 269–284, 2016. 1, 2, 5, 8

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 8

[12] Hyunjun Eun, Jinyoung Moon, Jongyoul Park, Chanho Jung, and Changick Kim. Learning to discriminate information for online action detection. In *CVPR*, pages 806–815, 2020. 2, 8

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6804–6815. IEEE, 2021. 2

[14] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 7, 8

[15] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. RED: reinforced encoder-decoder networks for action anticipation. In *BMVC*, 2017. 2, 7, 8

[16] Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, and Caiming Xiong. WOAD: weakly supervised online action detection in untrimmed videos. In *CVPR*, pages 1915–1923, 2021. 2, 8

[17] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021. 1, 2, 5, 7, 8

[18] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16081–16091. IEEE, 2022. 2

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18973–18990. IEEE, 2022. 8

[20] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017. 2

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New*

*Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[23] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 5

[24] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. 5

[25] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 2

[26] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 1

[27] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14024–14034, 2020. 1

[28] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *European Conference on Computer Vision*, pages 351–368, 2022. 2

[29] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *CoRR*, abs/2210.03929, 2022. 8

[30] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 2, 5, 8

[31] 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. 5, 8

[32] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 201–214. Springer, 2012. 1

[33] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015. 1

[34] Sumin Lee, Hyunjun Eun, Jinyoung Moon, Seokeon Choi, Yoonhyung Kim, Chanho Jung, and Changick Kim. Learning to discriminate information for online action detection: Analysis and application. *TPAMI*, pages 1–17, 2022. 5, 8

[35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 2

[36] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023. 2

[37] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. *arXiv preprint arXiv:2308.02236*, 2023. 2

[38] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. *CoRR*, abs/2206.01670, 2022. 8

[39] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *CoRR*, abs/2106.10271, 2021. 2

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 2

[41] Giovanni Pezzulo, Francesco Rigoli, and Karl Friston. Active inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology*, 134:17–35, 2015. 2

[42] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *CVPR*, 2018. 2, 5, 8

[43] Daniel L Schacter and Donna Rose Addis. The ghosts of past and future. *Nature*, 445(7123):27–27, 2007. 2

[44] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, volume 12361 of *Lecture Notes in Computer Science*, pages 154–171. Springer, 2020. 1, 7

[45] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1417–1426. IEEE Computer Society, 2017. 8

[46] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4669–4677, 2015. 1

[47] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *CoRR*, abs/2203.12602, 2022. 8

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021. 2

[49] 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. 2

[50] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 5

[51] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. KVT: k-nn attention for boosting vision transformers. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV*, volume 13684 of *Lecture Notes in Computer Science*, pages 285–302. Springer, 2022. 6

[52] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768, 2020. 2

[53] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 548–558. IEEE, 2021. 2

[54] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Long short-term transformer for online action detection. In *ICCV*, 2021. 2, 5, 7, 8

[55] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13577–13587. IEEE, 2022. 2

[57] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. Temporal recurrent networks for online action detection. In *ICCV*, pages 5532–5541, 2019. 2, 5, 7, 8

[58] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. In *NeurIPS*, 2021. 1, 2, 3, 5, 6, 7, 8

[59] Xinyu Xu, Yong-Lu Li, and Cewu Lu. Learning to anticipate future with dynamic context removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12734–12744, 2022. 7

[60] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16442–16453, June 2022. 2

[61] Le Yang, Junwei Han, and Dingwen Zhang. Colar: Effective and efficient online action detection by consulting exemplars. In *CVPR*, 2022. 2, 5, 8

[62] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 507–523. Springer, 2020. 1

[63] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *CoRR*, abs/2202.07925, 2022. 2

[64] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 6

[65] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *ICCV*, pages 13557–13567, 2021. 2

[66] 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. 8

[67] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 5, 6, 7, 8

[68] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. *CoRR*, abs/2212.04501, 2022. 8