Uncertainty-aware Action Decoupling Transformer for Action Anticipation

Hongji Guo[1]* Nakul Agarwal[2] Shao-Yuan Lo[2] Kwonjoon Lee[2] Qiang Ji[1]
[1]Rensselaer Polytechnic Institute, [2]Honda Research Institute, USA
{guohl1, jiq}@rpi.edu, {nakul_agarwal, shao-yuan_lo, kwonjoon_lee}@honda-ri.com

Abstract

*Human action anticipation aims at predicting what people will do in the future based on past observations. In this paper, we introduce **U**ncertainty-aware **A**ction **D**ecoupling **T**ransformer (UADT) for action anticipation. Unlike existing methods that directly predict action in a verb-noun pair format, we decouple the action anticipation task into verb and noun anticipations separately. The objective is to make the two decoupled tasks assist each other and eventually improve the action anticipation task. Specifically, we propose a two-stream Transformer-based architecture which is composed of a verb-to-noun model and a noun-to-verb model. The verb-to-noun model leverages the verb information to improve the noun prediction and the other way around. We extend the model in a probabilistic manner and quantify the predictive uncertainty of each decoupled task to select features. In this way, the noun prediction leverages the most informative and redundancy-free verb features and verb prediction works similarly. Finally, the two streams are combined dynamically based on their uncertainties to make the joint action anticipation. We demonstrate the efficacy of our method by achieving state-of-the-art performance on action anticipation benchmarks including EPIC-KITCHENS, EGTEA Gaze+, and 50-Salads.*

1. Introduction

Human action anticipation aims at predicting the future action before it happens based on the current observation. It is an important research topic for intelligent systems since it is widely applied for autonomous driving [44], human-robot interaction [31], and smart homes [19].

The task is very challenging as the future observation is unavailable and the anticipation needs to be made timely for real-time purposes [25]. Under most anticipation task settings [13, 28, 36, 38], the actions are represented as $(verb, noun)$ pairs, which means both verbs and nouns needed to be predicted correctly. Most existing methods [20, 25, 26, 40, 48, 61] for action anticipation tackle the task as an one-class action classification problem without considering the underlying dynamics and dependencies between verbs and nouns. These models directly output the action prediction, which is later decomposed into verb and noun predictions in a post-processing. However, this mechanism has a critical drawback. If either the verb or noun of the action is difficult to predict due to the limited visual cues, the action can be very difficult to predict correctly since it requires both verb and noun to be correct [59, 60]. On the other hand, if either verb or noun is known, the remaining part is much easier to predict. For example, the verb of "drinking coffee" can be easier predicted when knowing the "coffee" and the noun of "stretch dough" is easier to predict when knowing "stretch". In addition, the predictive uncertainty can be greatly reduced because the $p((verb, noun)|X)$ is converted to $p(verb|X, noun)$ or $p(noun|X, verb)$, where X is the input. The verb/noun information serves as a prior for the complementary part so the anticipation is simplified.

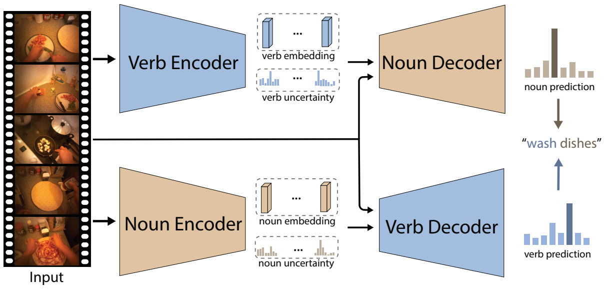


Figure 1. **An illustration of uncertainty-aware action decoupling transformer (UADT)**. UADT is composed of a verb-to-noun model (VtN) and a noun-to-verb model (NtV), which aims at anticipating noun and verb respectively. VtN anticipates the noun with the assistance of verb information and NtV anticipates the verb with the assistance of noun information. VtN and NtV are dynamically combined based on the predictive uncertainty.

To address the above issue of verb-noun modeling, we introduce *Uncertainty-aware Action Decoupling Transformer* (UADT), which decouples the action anticipation into verb and noun anticipations. Specifically, UADT is composed of a verb-to-noun model and a noun-to-verb

[*] Work done during internship at Honda Research Institute, USA.

model. Each model is composed of an encoder and a decoder. The encoder of the verb-to-noun model aims at generating verb embedding and its corresponding uncertainty. Then the embedding and its uncertainty are taken by the decoder to help the noun anticipation. We model the predictive uncertainty of the model because the encoder can generate bad embedding, which propagates the error to the following predictions. By quantifying the predictive uncertainty, we can leverage it to select reliable information and to filter the redundancy and irrelevance. In this way, the noun anticipation can be improved by benefiting from the verb information. Inversely, the noun-to-verb model first generates noun embeddings and the corresponding uncertainty. Then it performs the verb anticipation with assistance of noun information. In the end, we obtain the augmented noun and verb predictions. We dynamically combine them for the joint action anticipation based on their predictive uncertainties.

To train UADT, we adopt a two-stage training strategy. The encoders and decoders are trained with different loss functions to guide them for their specific purposes. We firstly train the encoders to generate the high-quality embeddings. Then we fix the encoders and train decoders for joint action anticipation. We evaluated UADT on both egocentric and third-person action anticipation datasets including EPIC-KITCHENS-100 [15], EGTEA Gaze+ [38], and 50-Salads [52]. Experiments results show the verb-to-noun model and noun-to-verb model effectively improve the anticipation of noun and verb respectively. We also demonstrate the effectiveness and benefits of proposed mechanisms and components by extensive ablation studies.

In summary, the main contributions of this paper are:

- We propose UADT for human action anticipation. By decoupling the action into verb and noun, we design and combine a verb-to-noun model and a noun-to-verb model. The two models assist each other to improve the joint action anticipation.
- By extending the model in a probabilistic manner, we quantify the predictive uncertainty, which is used for identifying informative embeddings and prediction fusion.
- UADT achieves state-of-the-art performance on egocentric and third-person action anticipation benchmark datasets including EPIC-KITCHENS-100, EGTEA Gaze+, and 50-Salads. In addition, extensive ablation studies demonstrate the effectiveness of the proposed mechanisms and components.

2. Related Work

2.1. Human Action Anticipation

Human action anticipation aims at predicting future actions before they occur. As for its practical applications, a number of benchmarks [15, 28, 36, 38, 52] have been built to boost related research. For anticipation, feature learning [22, 46, 55] and temporal modeling [3] are two main focuses. Recurrent neural network is widely adopted by many prior work [2–4, 20, 22, 35, 41, 51] to model the temporal relationship. For example, Furnari *et al*. [20] proposed rolling-unrolling LSTM (RULSTM) with a rolling LSTM to encode the historical information and an unrolling LSTM to decode the future actions.

Recently, transformer-based methods [25–27, 47, 57, 61] become the mainstream because of its strong capability for capturing long-range spatial-temporal dependencies. Typically, Girdhar *et al*. [26] proposed anticipative video transformer (AVT) with a pure self-attention design. Based on the transformer encoder, Girase *et al*. [25] introduced RAFTformer for real-time action forecasting with low inference latency. To leverage the overall goal of actions, multiple methods [43, 48, 49] are proposed to learn the hidden representations of goal to guide the anticipation. To utilize information from different sources such as audio and optical flow, various approaches [13, 20, 21, 32, 50, 65, 68] combined different modalities to improve the anticipation. Also, large language model (LLM) is also explored for action anticipation [66].

2.2. Uncertainty for Action Understanding

Uncertainty is a measure of prediction confidence [24, 33]. It not only represents the reliability of prediction, but also provides specific information about the model [42]. Uncertainty quantification techniques [33] have shown increasing importance in action understanding such as action recognition [30, 56, 64, 67] and temporal action detection [5, 7, 10–12, 29, 37, 62, 63]. Wang *et al*. [56] leverages uncertainty sampling for active learning to select most informative instances for action recognition. Guo *et al*. [30] proposed uncertainty-guided probabilistic transformer (UGPT) for complex action recognition. Specifically, uncertainty is quantified to train two models for low-uncertainty and high-uncertainty data respectively.

For anticipation task, a few work have explored the uncertainty of future actions. Furnari *et al*. [21] considered the uncertainty to design the loss function. Farha *et al*. [2] modeled the probability distribution of future action and generated multiple samples to account for the uncertainty. In some cases, the future action is almost impossible to infer, Suris *et al*. [53] proposed a hierarchical model to infer high-level activities when the simple action is difficult to anticipate. In this work, we model the uncertainty of verbs and nouns to identify reliable information and assist decision making.

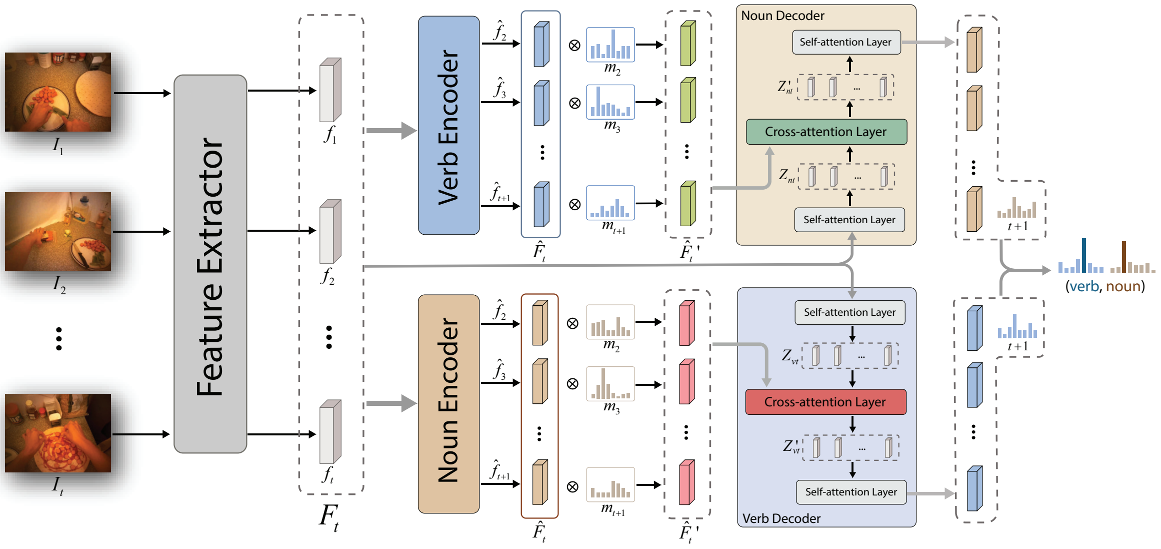


Figure 2. **Overall framework of UADT.** UADT is composed of a verb-to-noun (VtN) model on the top and a noun-to-verb (NtV) model on the bottom, which aims at anticipating the noun and the verb respectively. VtN is made up of a verb encoder and a noun decoder. NtV is made up of a noun encoder and a verb decoder. Given the input video, a pretrained backbone is firstly used to extract features. The extracted features are fed into two encoders to generate the verb/noun embeddings and their corresponding uncertainties, which are taken by the decoders to help the anticipation of the noun/verb. The decoders output the assisted verb and noun predictions. In the end, the VtN and NtV are dynamically combined based on their predictive uncertainties.

3. Method

In this section, we first give an overview of Uncertainty-aware Action Decoupling Transformer (UADT) in Sec. 3.1 and formulate the action anticipation task in Sec. 3.2. Then we introduce UADT's encoders and decoders in Sec. 3.3 and Sec. 3.4 respectively. The uncertainty-based fusion strategy is introduced in Sec. 3.5. Finally, we discuss the training procedures in Sec. 3.6.

3.1. Overview

An overall framework of UADT is shown in Figure 2. It is composed of a verb-to-noun (VtN) model and a noun-to-verb (NtV) model. Given the input video up to time t, features are firstly extracted by a pretrained backbone network. Then the extracted features are fed into a verb encoder and a noun encoder to generate the initial verb and noun embeddings. Meanwhile, the encoders also output the predictive uncertainty to identify informative and reliable embeddings. The initial verb embedding and its corresponding uncertainty are fed into noun decoder to assist the noun anticipation. And initial noun embedding and its corresponding uncertainty are fed into verb decoder to assist the verb anticipation. Besides noun and verb predictions, the decoders also output noun and verb uncertainties. Finally, the noun and verb predictions are dynamically combined based the uncertainty to make joint action anticipation.

3.2. Anticipation Problem Formulation

Human action anticipation aims at predicting future actions before they occur. In this work, we follow the settings of short-term action anticipation in [13, 14]. Mathematically, denote the input at time t as $\boldsymbol{X}_t = \{I_1, ..., I_t\}$, where $I_{t'}$ is the frame at time t'. The goal of anticipation is to predict the action in $(verb, noun)$ format at time $t + t_f$, where t_f is the time interval before the future action happens. So the problem can be formulated as a classification task as $y^*_{t+t_f} = \text{argmax}_{\hat{y}}\, p(\hat{y}_{t+t_f}|\boldsymbol{X}_t)$, where y denotes the action label. An illustration is shown in Figure 3.

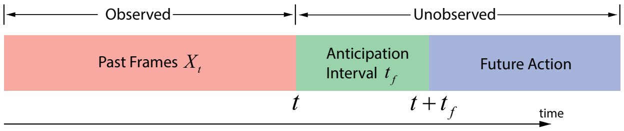


Figure 3. **Action anticipation illustration.** The task aims at predicting the future action after a time interval t_f given the observation up to time t.

3.3. UADT Encoders

UADT has a verb encoder (VE) and a noun encoder (NE). The objective of these encoders is to generate verb/noun embeddings that can be used as a prior to assist the anticipation of the complementary part. To achieve this, we build two probabilistic encoders based on Transformer [54] encoder to encode verb/noun information and quantify the predictive uncertainty. We quantify the uncertainty because not all the generated embeddings are reliable and useful. Uncertainty is used to identify informative embeddings that can better serve the decoders.

For the architecture, we follow the standard transformer encoder design. To capture the predictive uncertainty, we extend the model in a probabilistic manner. Specifically, we replace the feed-forward networks in the last encoder layer with Gaussian probabilistic layers [33] to model the parameter distribution. To learn the model, reparameterization trick [34] is used to perform forward pass and back-propagation. In this way, the model can generate multiple outputs based on the same input through sampling. Since only the forward process after the probabilistic layer needs to be repeated, the computational cost does not increase dramatically. Then the predictive uncertainty can be quantified based on these predictions.

To train the encoders to generate verb/noun-orientated embeddings, we design a verb/noun-guided training strategy. Given the input $\boldsymbol{X}_t$ at time t, a pretrained backbone first extracts the features as $\boldsymbol{F}_t = \{\boldsymbol{f}_1, ..., \boldsymbol{f}_t\}$. Then encoders make verb/noun anticipation at each time step based on $\boldsymbol{F}_t$:

$$\hat{\boldsymbol{v}}_2, ..., \hat{\boldsymbol{v}}_{t+1} = VE(\boldsymbol{F}_t),\ \hat{\boldsymbol{n}}_2, ..., \hat{\boldsymbol{n}}_{t+1} = NE(\boldsymbol{F}_t) \quad (1)$$

where $\boldsymbol{V}_t = \{\hat{\boldsymbol{v}}_2, ..., \hat{\boldsymbol{v}}_{t+1}\}$ and $\boldsymbol{N}_t = \{\hat{\boldsymbol{n}}_2, ..., \hat{\boldsymbol{n}}_{t+1}\}$ are predicted verbs and nouns.

To train the model for action anticipation, most approaches use cross-entropy-based loss functions to optimize the top-1 prediction. However, this may cause problems for our encoders since the top-1 prediction can be wrong and the error will propagate to the decoder part. To address this issue, we propose a top-K cross-entropy loss so that the encoder can tolerate certain erroneous predictions and encode more information. The loss function is defined as follows:

$$\mathcal{L}_{top\_K} = \begin{cases} -\log\sum_{k=1}^{K} p(\hat{y}_k), & \text{if top-}K\text{ prediction is correct} \\ -\sum_{c=1}^{C} \mathbb{1}(\hat{y}=c)\log p(c), & \text{o.w.} \end{cases} \quad (2)$$

where C is the total number of verb or noun classes, K is a hyper-parameter, and $\hat{y}_k$ denotes the top-k predicted label. From Eq. 2, a classification result has less penalty if its top-K predictions include the ground-truth verb/noun. In this way, the encoding space is extended to K verbs or nouns so it is more robust for wrong top-1 prediction. We empirically set $K = 5$ based on the ablation study in Figure 5a.

On the other hand, we also make the encoders do feature anticipation at each time step: $\hat{\boldsymbol{F}}_t = \{\hat{\boldsymbol{f}}_2, ..., \hat{\boldsymbol{f}}_{t+1}\}$, where $\hat{\boldsymbol{f}}_\tau$ is the predicted future feature of $\boldsymbol{f}_\tau$. Specifically, the output embeddings of the last encoder layer are treated as the anticipated features. We train them by minimizing the mean squared error loss $\mathcal{L}_{feat}$ between predicted features and true features in a self-supervised manner:

$$\mathcal{L}_{feat} = \|\boldsymbol{F}_{t+1} - \hat{\boldsymbol{F}}_t\|_2^2 \quad (3)$$

where $\boldsymbol{F}_{t+1} = \{\boldsymbol{f}_2, ..., \boldsymbol{f}_{t+1}\}$.

The VE and NE are trained separately by jointly minimizing the top-K verb/noun loss and the feature loss. The total encoder loss function can be written as:

$$\mathcal{L}_{en} = \mathcal{L}_{top\_K}^{verb/noun} + \lambda\mathcal{L}_{feat} \quad (4)$$

where λ is a hyper-parameter that measures the weight of the feature anticipation loss. After training the verb and noun encoders, their outputs encode the verb and noun information, which are utilized by the following decoders.

Uncertainty quantification. The generated embeddings above contain misleading information or redundancy. To address this issue, we measure the predictive uncertainty to identify the reliable embeddings. Modeling the uncertainty is effective for action anticipation because the observation of future action is unavailable and there is intra-class ambiguity. Even the same observed actions can lead to different future actions [53]. For example, “get cup” and “pour coffee” both exist in “make coffee” and “make tea”. And different people may perform the same action in different ways and in different orders. Therefore, it is important to model the predictive uncertainty for reliable future prediction.

Specifically, the predictive uncertainty is composed of epistemic uncertainty and aleatoric uncertainty [33]. Epistemic uncertainty, also known as model uncertainty, captures the lack of knowledge of model and is inversely proportional to the training data. In action anticipation task, epistemic uncertainty accounts for the unreliability of model for future actions. Aleatoric uncertainty, also known as data uncertainty, measures the noise in the data. This kind of uncertainty is related to the label and imperfectness of action data. Please refer to [1] for more details. These two types of uncertainties add up to the total predictive uncertainty. As epistemic uncertainty account for the internal property of model for the unknowns, it is more effective for anticipation task. We demonstrate this claim in the ablation study (§ 4.4). In this work, we mainly leverage the epistemic uncertainty.

By extending the model in a probabilistic manner, we learned the distribution of parameters. So we can obtain N sets of parameters $\{\theta_1, ..., \theta_N\}$ by sampling and then get N predictions from the same input by repeating the forward process with different parameters. Generally, the total predictive uncertainty is quantified as the entropy of the predictions as $\mathcal{H}[\hat{y}|x] = -\sum_{c=1}^{C} p(\hat{y}|x)\log p(\hat{y}|x)$, where y is the output label and x is the input. The epistemic uncertainty can be quantified as follows [42]:

$$\mathcal{U}_e \approx \mathcal{H}[\frac{1}{N}\sum_{n}^{N} p(\hat{y}|x,\theta_n)] - \frac{1}{N}\sum_{n}^{N}\mathcal{H}[p(\hat{y}|x,\theta_n)] \quad (5)$$

In Eq. 5, we use average of N samples to approximate the true value since it is intractable to integrate over the parameter space. The second term on the right is an approximation of aleatoric uncertainty:

$$\mathcal{U}_a \approx \frac{1}{N}\sum_{n}^{N} \mathcal{H}[p(\hat{y}|x, \theta_n)] \tag{6}$$

We follow the same procedure for every time step so that every generated embedding has its corresponding uncertainty. The generated embeddings along with their uncertainties are fed into the decoder make the final anticipation.

3.4. UADT Decoders

The objective of decoders is to make the verb/noun anticipation by leveraging the noun/verb embeddings and uncertainties from the encoders. They take both the extracted features from backbone and the embeddings from the encoders.

As shown in Figure 2, the decoders are composed of self-attention layers and cross-attention layers. The self-attention layers are transformer encoders with causal masks to make sure each step can only access its past information. They first take the input features $\boldsymbol{F}_t$ to generate the intermediate noun embeddings $\boldsymbol{Z}_{nt} = \{\boldsymbol{z}_{n1}, ..., \boldsymbol{z}_{nt}\}$ and intermediate verb embeddings $\boldsymbol{Z}_{vt} = \{\boldsymbol{z}_{v1}, ..., \boldsymbol{z}_{vt}\}$. $\boldsymbol{Z}_{nt}$ and $\boldsymbol{Z}_{vt}$ are orientated to anticipate the noun and verb respectively, which are used for cross-attention.

Before the encoder embeddings enter the cross-attention layer, we apply an uncertainty mask $\boldsymbol{M}_t = \{\boldsymbol{m}_2, ..., \boldsymbol{m}_{t+1}\}$ to select the most informative embedding. We assume the embeddings with large uncertainty tend to be less reliable and less relevant to the actions being anticipated. So the weights of masks are inversely proportional to the uncertainty. Specifically, the weights of the uncertainty mask at time t can be computed as follows:

$$\boldsymbol{m}_{t'} = 1 - (\mathcal{U}_t - \mathcal{U}_{min})/(\mathcal{U}_{max} - \mathcal{U}_{min}) \tag{7}$$

where $\mathcal{U}_{t'}$ is the epistemic uncertainty of embedding at time t', and $\mathcal{U}_{max}$, $\mathcal{U}_{min}$ are the maximum and minimum epistemic uncertainty within each batch. The uncertainty mask is multiplied to the encoder embeddings at each time step to generate the weighted embeddings $\hat{\boldsymbol{F}'}_t = \{\boldsymbol{m}_2\hat{\boldsymbol{f}}_2, ..., \boldsymbol{m}_{t+1}\hat{\boldsymbol{f}}_{t+1}\}$. Same procedures are used for both models. Then we perform the cross attention between $\boldsymbol{Z}_{vt}$/$\boldsymbol{Z}_{nt}$ and encoder embeddings $\hat{\boldsymbol{F}'}_t$ of noun/verb model. After cross-attention, the updated embeddings are fed into the last self-attention layer to generate the final embeddings.

To train the decoders, we first have a standard cross-entropy loss to train the decoder at time t for next verb/noun anticipation:

$$\mathcal{L}_{next} = -\log \hat{y}_t[c_{t+1}] \tag{8}$$

where $\hat{y}_t$ is the predicted label at time t and c_{t+1} is the ground-truth label of frame $t+1$.

In addition, we follow the same procedures as the encodes to make feature anticipation by minimizing $\mathcal{L}_{feat}$ in Eq. 3. We also train the model to do verb/noun anticipation before time t. Specifically, each output embedding goes through a linear layer to output the verb/action prediction. We train this sub-task with a cross-entropy loss as follows:

$$\mathcal{L}_{verb/noun} = -\sum_{\tau=1}^{t-1} \log \hat{y}_\tau[c_{\tau+1}] \tag{9}$$

where $\hat{y}_\tau$ is the predicted verb or noun label at time τ.

The total loss function for training the decoders can be written as:

$$\mathcal{L}_{de} = \mathcal{L}_{next} + \lambda_1\mathcal{L}_{feat} + \lambda_2\mathcal{L}_{verb/noun} \tag{10}$$

where λ_1 and λ_2 are hyper-parameters.

Decoder uncertainty. The last self-attention layer in the decoder is also extended in a probabilistic manner. And decoders output the predictive uncertainty of the anticipated verb and noun. The noun and verb uncertainties represent the the reliability of the verb-to-noun model and noun-to-verb model respectively. The noun and verb predictions along with their uncertainties are combined for the final action anticipation.

3.5. Uncertainty-based Fusion

Joint action anticipation. To combine the predictions of verb-to-noun model and noun-to-verb model, we proposed an uncertainty-based fusion strategy. We assume the prediction with low uncertainty is more reliable so it should be assigned higher weights. The fusion of the two models can be written as:

$$\begin{aligned} p(verb) &= \alpha p^{en}_{v\to n} + (1-\alpha)p^{de}_{n\to v} \\ \alpha &= \sigma((\mathcal{U}_{n\to v} - \mathcal{U}^{v}_{min}/\mathcal{U}^{v}_{max} - \mathcal{U}^{v}_{min})) \end{aligned} \tag{11}$$

$$\begin{aligned} p(noun) &= \beta p^{en}_{n\to v} + (1-\beta)p^{de}_{v\to n} \\ \beta &= \sigma((\mathcal{U}_{v\to n} - \mathcal{U}^{n}_{min}/\mathcal{U}^{n}_{max} - \mathcal{U}^{n}_{min})) \end{aligned} \tag{12}$$

where p denotes the prediction and σ is the sigmoid function. α and β are functions of the predictive epistemic uncertainty. In this way, the prediction that has high uncertainty is less considered in the final anticipation. The fusion is dynamic since it depends on the input uncertainty. By considering the uncertainty of future verb/noun in the decision process, the final anticipation is made by the most reliable verb and noun combination.

Post-processing. After the fusion of verb-to-noun model and noun-to-verb model, we obtain the joint action prediction. However, the verbs and nouns are predicted separately, which means some $(verb, noun)$ pairs can be implausible such as "drinking potatoes". To correct implausible verb-noun pairs, we perform a post-processing by selecting the verb-noun pair that has the maximum joint probability among valid $(verb, noun)$ combinations.

$$(verb, noun)^* = \underset{(verb,noun)\in\mathcal{Y}}{\operatorname{argmax}} \; p(verb)p(noun) \tag{13}$$

Algorithm 1 UADT Training

```
Input: $\mathcal{D} = \{\boldsymbol{X}_n \in \mathbb{R}^{T\times H\times W\times C}, \boldsymbol{Y}_n \in \mathbb{R}^T\}_{n=1}^N$ - data
Output: Encoder and decoder parameters $\{\theta_{en}, \theta_{de}\}$
  Extract $\boldsymbol{F}_t$ from $\boldsymbol{X}_t$ by a backbone
  Training encoders
  Make predictions $\hat{\boldsymbol{V}}_t$ and $\hat{\boldsymbol{N}}_t$
  Compute anticipated features $\hat{\boldsymbol{F}}_t$
  Optimizing $\theta_{en}$ by minimizing $\mathcal{L}_{en}$ in Eq. 4
  Compute encoder uncertainty $\mathcal{U}_e$ by Eq. 5
  Training decoders
  Generate intermediate embeddings $\boldsymbol{Z}_{nt}$ and $\boldsymbol{Z}_{vt}$
  Cross-attention between $\boldsymbol{Z}_{nt}/\boldsymbol{Z}_{vt}$ and $\hat{\boldsymbol{F}'}_t$
  Predict $\boldsymbol{V}_t$ and $\boldsymbol{N}_t$
  Compute anticipated features $\hat{\boldsymbol{F}}_t$
  Optimize $\theta_{de}$ by minimizing $\mathcal{L}_{de}$ in Eq. 10
  return $\theta = \{\theta_{en}, \theta_{de}\}$
```

where $\mathcal{Y}$ is the set that contains all plausible $(verb, noun)$ combinations.

3.6. Training Procedures

To train the UADT, we adopt a two-stage (2S) training strategy. We first train the verb encoder and noun encoder separately by minimizing $\mathcal{L}_{en}$ in Eq. 4. Afterwards, we fix the encoders and train the noun decoder and verb decoder by minimizing $\mathcal{L}_{de}$ in Eq. 10. The training procedures are summarized in Algorithm 1. In addition, we also trained the model in an end-to-end (E2E) manner. E2E training gives better performance than the 2S training. An ablation study of training strategies is available in Sec. 4.4.

4. Experiments

We first introduce the benchmark and evaluation metrics (§ 4.1). Then we provide the implementation details (§ 4.2). Next we compare UADT with state-of-the-art methods (§ 4.3). Finally, we present ablation studies of the proposed mechanisms and components (§ 4.4).

4.1. Datasets and Evaluation Metrics

EPIC-KITCHENS-100 (EK100) [15] is a large-scale egocentric video dataset. It contains 700 cooking activity videos. There are 3806 actions with 97 verbs and 300 nouns. In this work, we evaluate our proposed method on the validation dataset as previous work [20, 25] without additional training data. Following the settings in [21], we report the top-5 recalls of action, verb, and noun.

EGTEA GAZE+ [38] is a large-scale dataset for first-person-view (FPV) actions and gaze. It contains 28 hours cooking activity videos from 86 unique sessions of 32 subjects. There are totally 106 actions with 19 verbs and 51 nouns. Top-1 accuracy is used as the evaluation metric.

Method	Init	Modality	Top-5 Recall		
			Verb	Noun	Action
TempAgg [50]	IN1k	RGB	24.2	29.8	13.0
RULSTM [20]	IN1k	RGB	-	-	13.3
RULSTM [20]	IN1k	RGB+Flow+Obj	30.8	27.8	14.0
TempAgg [50]	IN1k	Flow+Obj+ROI	21.2	31.4	14.7
AVT [26]	IN21k	RGB	30.2	31.7	14.9
AVT+ [26]	IN21k	RGB+Obj	28.2	32.0	15.9
TSN-AVT+ [26]	IN21k	RGB+Obj	31.8	25.5	14.8
MeMViT [61]	K400	RGB	32.8	33.2	15.1
RAFTformer [25]	K400+IN1k	RGB	33.3	35.5	17.6
UADT (ours)	K400	RGB	**35.2**	**38.5**	**18.8**
MeMViT [61]	K700	RGB	32.2	37.0	17.7
RAFTformer [25]	K700	RGB	33.7	37.1	18.0
RAFTformer-2B [25]	K700+IN1k	RGB	33.8	37.9	19.1
UADT (ours)	K700	RGB	**38.2**	**41.4**	**20.3**
UADT (ours)	K700	RGB+Flow+Obj	**43.5**	**46.6**	**23.0**

Table 1. **Experiment results on EK100 validation set.** UADT achieves state-of-the-art performance under different settings.

50-Salads [52] is a third-person video dataset for action understanding. It captures 25 people preparing 2 mixed salads in 966 activity instances. There are totally 17 different actions. Following [50], we report the top-1 action accuracy over the pre-defined splits for comparison.

4.2. Implementation Details

Feature extraction. For EK100 dataset, we adopt MViT-b [17, 39] as the backbone. We pre-trained the 16×4 MViT-b on Kinetics-400 [8] for action classification. The 16×4 model uses 16 frames sampled 4 frames apart at 30fps, which leads to 2 seconds for each clip at 8fps. On the other hand, the Kinetics-700 [9] pretrained features are obtained by a 32×3 MViT, which uses 32 frames sampled 3 frames apart at 30fps. For EGTEA Gaze+, a TSN [58] pretrained on ImageNet-1K is used to extract features following the procedure in RULSTM [20]. For 50-Salads dataset, we used the I3D [8] features provided in [18]. In this way, we use the same feature as the SOTA [48] for fair comparisons. More details can be found in the supplementary.

Settings. The proposed framework is implemented in PyTorch [45]. The UADT is optimized using AdamW optimizer with momentum 0.8 and weight decay of 10^{-3}. We train the model for 50 epochs using a cosine scheduler with a 20 warmup epochs. The batch size is set to 512. The base learning rate is set to 10^{-4} and end learning rate is set to 10^{-6}. The batch size is set to 512. The dropout rate of transformer is set to 0.25. λ is set to 6 in $\mathcal{L}_{en}$. And we set $\lambda_1 = 5$ and $\lambda_2 = 0.1$ in $\mathcal{L}_{de}$. More details are available in the supplementary.

4.3. Comparison to state-of-the-art

EK100. The experiment results on EK100 are shown in Table 1. Only using the RGB modality, our proposed UADT outperforms the state-of-the-art RAFTformer [25] by a large margin with both K400 and K700 features. We

Method	Init	Modality	Top-5 Recall
DMR [55]	-	RGB	38.1
ASTN [13]	TSN/IN1k	RGB+Flow	31.6
MCE [21]	TSN/IN1k	RGB+Flow	43.8
TCN [6]	-	RGB	47.1
FN [16]	VGG-16	RGB	42.7
RED [23]	VGG-16/TS	RGB+Flow	54.6
RULSTM [20]	TSN/IN1k	RGB+Flow+Obj	58.6
RAFTformer [25]	TSN/IN1k	RGB	63.5
UADT (ours)	TSN/IN1k	RGB	**68.4**

Table 2. **Experiment results on EGTEA Gaze+.** UADT achieves SOTA performance using the same TSN/IN1k features.

Method	Top-1 Acc. (%)
DMR [55]	6.2
RNN [3]	30.1
CNN [3]	29.8
ActionBanks [50]	40.7
AVT [26]	48.0
RAFTformer* [25]	53.2
Latent-goal [48]	59.6
UADT (ours)	**62.7**

Table 3. **Experiment results on 50-Salads.** Using the same I3D features as prior work, UADT outperforms all state-of-the-art methods. * indicates reproduced results.

RGB	Flow	Obj	Top-5 Recall (%)		
			Verb	Noun	Action
✓			38.2	41.4	20.3
✓	✓		41.7	42.9	21.2
✓		✓	41.0	44.6	21.5
✓	✓	✓	**43.5**	**46.6**	**23.0**

Table 4. **Experiment results of different modalities on EK100 val with K700 features.** By incorporating additional modalities, the performance is significantly improved.

Method	K400 (Top-5 recall)			K700 (Top-5 recall)		
	Verb	Noun	Action	Verb	Noun	Action
VtN-b	-	35.5	-	-	38.7	-
VtN-U	-	**37.7**	-	-	**40.8**	-
NtV-b	31.7	-	-	35.1	-	-
NtV-U	**34.3**	-	-	**37.5**	-	-
UADT-b	33.2	36.5	18.0	36.0	39.1	19.4
UADT	**35.2**	**38.5**	**18.8**	**38.2**	**41.4**	**20.3**

Table 5. **Ablation study of encoder uncertainty on EK100 val.** “b” denotes the baseline version and “U” denotes the uncertainty-based version.

Method	K400 (Top-5 recall)			K700 (Top-5 recall)		
	Verb	Noun	Action	Verb	Noun	Action
Baseline	33.2	36.5	18.0	36.0	39.1	19.4
Total-U	34.7	38.0	18.5	37.7	41.0	20.1
Aleatoric-U	33.6	37.1	18.2	36.8	40.2	19.6
Epistemic-U	**35.2**	**38.5**	**18.8**	**38.2**	**41.4**	**20.3**

Table 6. **Ablation study of different types of uncertainties on EK100 val.** Epistemic uncertainty is the most effective.

also show that the performance can be further boosted by incorporating more modalities in the ablation study (§ 4.4).

EGTEA Gaze+. The comparison is shown in Table 2. Using the same TSN/1N1k features, our UADT significantly outperforms the RAFTformer by 4.9%.

50-Salads. The results are shown in Table 3. Our UADT outperforms the Latent-goal [48] by 3.1% using the same I3D features. This demonstrates that UADT can generalize to third-person dataset for anticipation.

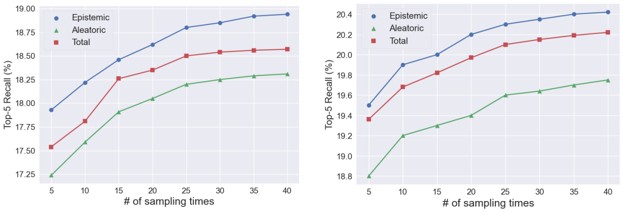
4.4. Ablation Studies

Different input modalities. To further study UADT, we incorporate results using additional modalities including the optical flow and object features. Specifically, we concatenate the feature vectors of different modalities at each time step. The results are shown in Table 4. By comparison, the performance is significantly improved by adding extra modalities. The optical flow features significantly improve the verb anticipation since they contain motion patterns. The object features are relatively effective for noun anticipation as the detected objects are highly-related to the noun of the action.

Encoder uncertainty modeling. To demonstrate the effectiveness of uncertainty from the decoder, we implemented a baseline verb-to-noun model (VtN-b) and noun-to-verb model (NtV-b) without uncertainty mask. The two models have exactly the same architectures as the probabilistic ones we proposed. A comparison of performance is shown in Table 5. From the results, the uncertainty-based single-stream VtN/NtV and UADT outperform their baseline version, which demonstrates the effectiveness of the encoder uncertainty modeling.

Types of uncertainty. We quantify the epistemic uncertainty in Eq. 5, aleatoric uncertainty in Eq. 6, as well as the total uncertainty. Then we generate the uncertainty masks based on different types of uncertainties and test the model. The comparison is shown in Table 6. The baseline method is implemented in same architecture without uncertainty modeling. From the results, the epistemic uncertainty is more effective than the other two types of uncertainties, which demonstrates our claim in Sec. 3.3.

Uncertainty sampling. In the uncertainty quantification process, we repeat the forward process to obtain N predictions. The number of samples affects the accuracy of uncertainty and further affects the anticipation performance. We varied the number of samples for different types of uncertainties. A comparison is shown in Figure 4. From the plots, it takes around 25 sampling times to obtain the relatively stable performance. Although the performance is still improving by increasing sampling times, we reported the performance of 25 sampling times in this paper due to the efficiency concern. The inference latency comparison with



(a) K400 features (b) K700 features

Figure 4. **Ablation study of the uncertainty sampling and different types of uncertainty on EK100 val.** By increasing the number of samples, the uncertainty quantification is more accurate and it further improves the anticipation.

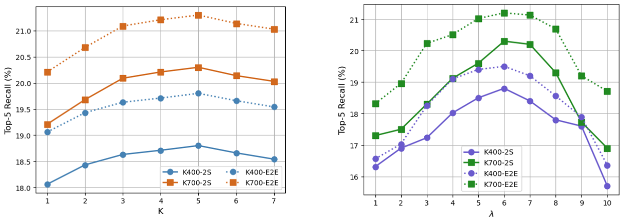
Training	Feature	Top-5 Recall		
		Verb	Noun	Action
Two-stage	K400	35.2	38.5	18.8
E2E-one-stage		37.3	40.1	19.3
E2E-two-stage		**37.4**	**40.4**	**19.5**
Two-stage	K700	38.2	41.4	20.3
E2E-one-stage		41.2	42.8	21.1
E2E-two-stage		**41.5**	**43.0**	**21.2**

Table 7. **Ablation study of training strategies on EK100 val.** The proposed two-stage training can be improved by E2E training.

different sampling times can be found in the supplementary.

Training strategies. For UADT, we adopt a two-stage (2S) training mechanism. The encoders are fixed after the first-stage training and decoders are trained afterwards. To better optimize the model for anticipation, we also trained the model in an end-to-end (E2E) manner. We implemented two types of E2E training, namely the one-stage version and two-stage version. Specifically, the one-stage E2E trains the encoders and decoders together from scratch. For the two-stage E2E, we train the encoders first by minimizing $\mathcal{L}_{en}$. Then we jointly train the encoders and decoders in the second stage by minimizing $\mathcal{L}_{de}$. The experiment results are shown in Table 7. The end-to-end methods obtain better results because the encoders are further optimized for anticipation after being trained for generating embeddings. The two-stage E2E converges faster than the one-stage E2E since the encoders are learned beforehand. In the comparison with state-of-the-art methods, we reported the results obtained by the 2S training instead of E2E training because the latter increases the training cost. A detailed comparison of training cost is available in supplementary.

Loss function. The encoder loss function $\mathcal{L}_{en}$ is composed of a top-K verb/noun loss and mean-squared error feature loss. To study the effect of K and the balance between two terms. We varied K and λ during training. The results with different K on EK100 val are shown in Figure 5a. Note the top-K loss becomes standard cross-entropy loss when $K = 1$. So the comparison also demonstrates the superiority of the top-K loss against the standard cross-entropy loss. The ablation study of λ is plotted in Figure 5b. From the results, we empirically set $K = 5$ and $\lambda = 6$ since these settings output best performance under different settings. The ablation studies of λ_1 and λ_2 in the decoder loss function can be found in the supplementary.



(a) Top-K loss (b) Feature loss

Figure 5. **Ablation study of encoder loss function on EK100 val.** The top-K loss effectively improve the performance. We set $K = 5$ and $\lambda = 6$ since they output the best results under different settings.

Method	K400 (Top-5 recall)			K700 (Top-5 recall)		
	Verb	Noun	Action	Verb	Noun	Action
VtN	-	37.7	-	-	40.8	-
NtV	34.3	-	-	37.5	-	-
Early fusion	32.1	35.9	17.9	36.1	38.9	19.2
Late fusion	34.5	37.8	18.3	37.7	40.9	20.0
Attention [20]	34.8	38.0	18.5	37.8	41.1	20.1
Uncertainty (§ 3.5)	**35.2**	**38.5**	**18.8**	**38.2**	**41.4**	**20.3**

Table 8. **Ablation study of fusion strategies on EK100 val.** All methods use by two-stage training with epistemic uncertainty.

Comparison of fusion strategies. In this work, we proposed an uncertainty-based fusion strategy of the verb-to-noun model and noun-to-verb model. To demonstrate its effectiveness, we compare it with other types of fusion methods. First, we test the early fusion method by combing the predictions of verb and noun encoders. Second, we test late fusion by combing the predictions of both verb-to-noun model and noun-to-verb model. Additionally, we test the attention fusion method proposed in [20]. The results and comparison are shown in Table 8. The uncertainty-based fusion outperforms other methods using either K400 or K700 features, which demonstrates its effectiveness.

5. Conclusion and Future Work

In this paper, we introduced UADT for action anticipation. By combining a verb-to-noun model and a noun-to-verb model, the verb and noun predictions assist each other to improve joint action anticipation.

In the future, we plan to extend it for long-term action anticipation that aims at predicting a larger number of future actions. And we also plan to leverage large language models to capture the verb and noun dependencies.

Acknowledgment: The work is supported in part by the National Science Foundation award IIS 2236026 and by the financial support from Rensselaer Polytechnic Institute.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021. 4

[2] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[3] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 2, 7

[4] Yazan Abu Farha, Qiuhong Ke, Bernt Schiele, and Juergen Gall. Long-term anticipation of activities with cycle consistency. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42*, pages 159–173. Springer, 2021. 2

[5] Anurag Arnab, Chen Sun, Arsha Nagrani, and Cordelia Schmid. Uncertainty-aware weakly supervised action detection from untrimmed videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 751–768. Springer, 2020. 2

[6] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 7

[7] Wentao Bao, Qi Yu, and Yu Kong. Opental: Towards open set temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2979–2989, 2022. 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[9] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 6

[10] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 192–208. Springer, 2022. 2

[11] Yunze Chen, Mengjuan Chen, Rui Wu, Jiagang Zhu, Zheng Zhu, Qingyi Gu, and Horizon Robotics. Refinement of boundary regression using uncertainty in temporal action localization. In *BMVC*, page 5, 2020.

[12] Yunze Chen, Mengjuan Chen, and Qingyi Gu. Class-wise boundary regression by uncertainty in temporal action detection. *IET Image Processing*, 16(14):3854–3862, 2022. 2

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 1, 2, 3, 7

[14] 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. *arXiv preprint arXiv:2006.13256*, 2020. 3

[15] 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*, pages 1–23, 2022. 2, 6

[16] Roeland De Geest and Tinne Tuytelaars. Modeling temporal structure with lstm for online action detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1549–1557. IEEE, 2018. 7

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 6

[18] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 6

[19] Iram Fatima, Muhammad Fahim, Young-Koo Lee, and Sungyoung Lee. A unified framework for activity recognition-based behavior analysis and action prediction in smart homes. *Sensors*, 13(2):2682–2699, 2013. 1

[20] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019. 1, 2, 6, 7, 8

[21] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 6, 7

[22] 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

[23] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 7

[24] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. 2

[25] Harshayu Girase, Nakul Agarwal, Chiho Choi, and Karttikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 18759–18769, 2023. 1, 2, 6, 7

[26] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021. 1, 2, 6, 7

[27] 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

[28] 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. 1, 2

[29] Hongji Guo, Zhou Ren, Yi Wu, Gang Hua, and Qiang Ji. Uncertainty-based spatial-temporal attention for online action detection. In *European Conference on Computer Vision*, pages 69–86. Springer, 2022. 2

[30] Hongji Guo, Hanjing Wang, and Qiang Ji. Uncertainty-guided probabilistic transformer for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20052–20061, 2022. 2

[31] Kelsey P Hawkins, Nam Vo, Shray Bansal, and Aaron F Bobick. Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 499–506. IEEE, 2013. 1

[32] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3125. IEEE, 2016. 2

[33] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2, 4

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[35] Yu Kong, Shangqian Gao, Bin Sun, and Yun Fu. Action prediction from videos via memorizing hard-to-predict samples. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[36] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 1, 2

[37] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1854–1862, 2021. 2

[38] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 1, 2, 6

[39] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 6

[40] Tianshan Liu and Kin-Man Lam. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13904–13913, 2022. 1

[41] Siyuan Brandon Loh, Debaditya Roy, and Basura Fernando. Long-term action forecasting using multi-headed attention-based variational recurrent neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2419–2427, 2022. 2

[42] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018. 2, 4

[43] 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

[44] Angelos Mavrogiannis, Rohan Chandra, and Dinesh Manocha. B-gap: Behavior-guided action prediction for autonomous navigation. *arXiv preprint arXiv:2011.03748*, 1 (2), 2020. 1

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[46] Cristian Rodriguez, Basura Fernando, and Hongdong Li. Action anticipation by predicting future dynamic images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

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

[48] Debaditya Roy and Basura Fernando. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2745–2753, 2022. 1, 2, 6, 7

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

[50] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 154–171. Springer, 2020. 2, 6, 7

[51] 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

[52] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013. 2, 6

[53] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12617, 2021. 2, 4

[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. 3

[55] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016. 2, 7

[56] Hanmo Wang, Xiaojun Chang, Lei Shi, Yi Yang, and Yi-Dong Shen. Uncertainty sampling for action recognition via maximizing expected average precision. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018. 2

[57] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13824–13835, 2023. 2

[58] 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 *European conference on computer vision*, pages 20–36. Springer, 2016. 6

[59] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12249–12256, 2020. 1

[60] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 1

[61] 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. 1, 2, 6

[62] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4146–4155, 2021. 2

[63] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised and unsupervised temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5252–5267, 2022. 2

[64] Yuanhao Zhai, Ziyi Liu, Zhenyu Wu, Yi Wu, Chunluan Zhou, David Doermann, Junsong Yuan, and Gang Hua. Soar: Scene-debiasing open-set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10244–10254, 2023. 2

[65] 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*, pages 1316–1322, 2021. 2

[66] Qi Zhao, Ce Zhang, Shijie Wang, Changcheng Fu, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023. 2

[67] Rui Zhao, Wanru Xu, Hui Su, and Qiang Ji. Bayesian hierarchical dynamic model for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7733–7742, 2019. 2

[68] 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