Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation

Zeyun Zhong[1,2*], David Schneider[2*], Michael Voit[1], Rainer Stiefelhagen[2], and Jürgen Beyerer[1,2]

[1]Fraunhofer IOSB, Karlsruhe {firstname.lastname}@iosb.fraunhofer.de
[2]Karlsruhe Institute of Technology (KIT) {firstname.lastname}@kit.edu

Abstract

Although human action anticipation is a task which is inherently multi-modal, state-of-the-art methods on well known action anticipation datasets leverage this data by applying ensemble methods and averaging scores of uni-modal anticipation networks. In this work we introduce transformer based modality fusion techniques, which unify multi-modal data at an early stage. Our Anticipative Feature Fusion Transformer (AFFT) proves to be superior to popular score fusion approaches and presents state-of-the-art results outperforming previous methods on EpicKitchens-100 and EGTEA Gaze+. Our model is easily extensible and allows for adding new modalities without architectural changes. Consequently, we extracted audio features on EpicKitchens-100 which we add to the set of commonly used features in the community. [1]

1. Introduction

Beyond human action recognition, anticipating possible future actions, as displayed in Figure 1, is one of the most important tasks for human machine cooperation and robotic assistance, e.g. to offer a hand at the right time or to generate proactive dialog to provide more natural interactions. As the anticipation results are just assumptions, this tends to be significantly more challenging than traditional action recognition, which performs well with today's well-honed discriminative models [13, 35]. As modeling long temporal context is often crucial for anticipation [16, 42, 20], many such methods were proposed in recent years, including clustering [21, 36], attention [42] and recurrence [16]. While vision based systems are the de-facto standard for action anticipation [16, 20, 48], additionally using other supporting modalities like optical flow features [47, 7, 28] or

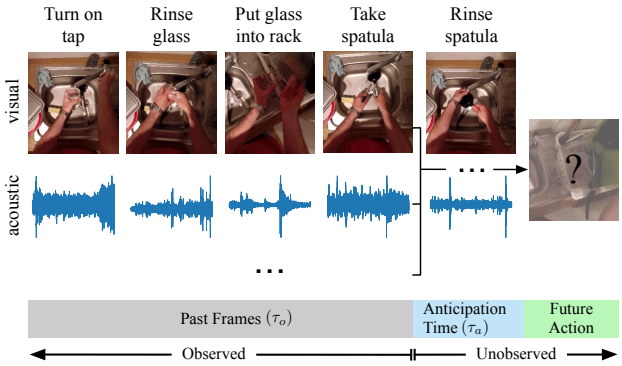

Figure 1: The action anticipation task aims to use the observed video segment of length τ_o to anticipate a future action τ_a seconds before it happens.

knowledge about objects in the scene [15] have shown to be beneficial. In recent work [28, 27, 37], audio has been explored and shown to be complementary with appearance for action recognition in first-person vision. Consistent with most multi-modal action recognition models [47, 7], anticipation models typically use score fusion (i.e., averaging predictions computed based on each single modality) to fuse different modalities. While averaging using fixed weights, including simple averaging [42] and weighted averaging [20], shows already superior results over the uni-modal baseline, Furnari et al. [16] show that assigning each modality with dynamical importance for the final prediction is particularly beneficial for anticipating egocentric actions. Inspired by the classical view of multisensory integration, i.e., information across the senses gets merged after the initial sensory processing is completed [6, 45], we take the mid-level fusion strategy in this work. We present a transformer-based feature fusion model, Anticipative Feature Fusion Transformer (AFFT), which successfully combines multi-modal features in a mid-level fusion process where features are first fused and the fused representations

*Equal contribution
[1]Code: https://github.com/zeyun-zhong/AFFT

are utilized to anticipate next actions, different from all late and score fusion methods mentioned above. Our method is based on features and does not require end-to-end training of feature extractors. We see this as a major advantage since recent state-of-the-art results on various tasks have been driven by large foundation models which are difficult and resource intensive to train. By combining strong feature extractors like OMNIVORE [22] with mid-level feature fusion, we achieve state-of-the art results on common action anticipation datasets without the need for fine-tuning them.

In summary, our main contributions are:

- The **A**nticipative **F**eature **F**usion **T**ransformer (AFFT), which successfully performs mid-level fusion on extracted features, improves significantly over score fusion based approaches and provides state-of-the-art performance on EpicKitchens-100 action anticipation and competing results on EGTEA Gaze+;
- A comparison of multiple self-attention and cross-attention based feature fusion strategies as well as detailed hyper parameter ablations for our final model;
- Extracted audio and OMNIVORE-based RGB features of EpicKitchen-100 which we provide to the community and an analysis of temporal and modality-wise performance contributions and model attention values.

2. Related Work

Action anticipation aims to predict future actions given a video clip of the past and present. While many approaches investigated different forms of action and activity anticipation from third person video [17, 12, 29, 23], the first-person (egocentric) vision has recently gained popularity along with development of multiple challenge benchmarks to support it [8, 9, 32]. To model the temporal progression of past actions, [16] proposed using an LSTM to summarize the past and another LSTM for future prediction. [42] made use of long-range past information and used an adapted version of the attention mechanism to aggregate short-term ('recent') and long-term ('spanning') features. To maintain the sequential temporal evolution while addressing the problem of modeling long-range temporal dependencies of recurrent architectures, a variation of GPT-2 [40] has been recently proposed in [20]. We propose a transformer based feature fusion model to effectively fuse multiple modalities, and follow [20] to use a generative language model for future action prediction.

Multi-modal fusion for action anticipation. The modalities typically used in prior work for egocentric vision are RGB, objects and optical flow [16, 42, 49, 50, 20]. To fuse information contained in different modalities, anticipation models typically utilize a late fusion strategy, similar to many multi-modal action recognition models [47, 7, 28]. These fusion methods can be broadly divided into score fusion and feature fusion. While in score fusion, the predicted future action scores of each modality are combined using either fixed weights, in form of simple averaging [42, 49] or weighted averaging [20], or dynamic weights based on the scene [16], the feature fusion combines the predicted future action feature and an additional feed-forward layer is utilized to generate the action score [50]. Different from the late fusion strategy, we take the mid-level fusion strategy inspired by the classical view of multisensory integration [6, 45]. Specifically, we adopt the multi-head attention mechanism [46] to combine different modalities at each timestamp and utilize the variation of GPT-2 following [20] to analyze the temporal evolution of the fused past features and predict future action features. Finally, a feed-forward layer is used to predict the future action class.

Audio-visual learning. Recent work used audio for an array of video understanding tasks, including self-supervised representation learning [5, 3, 30], audio-visual source separation [38, 2, 11], localizing sounds in video frames [4, 43], generating sounds from video [39, 52, 18], leveraging audio for efficient action recognition [31, 19], and utilizing audio to improve classification performance of action recognition [28, 27, 37]. Different from all the work above, we focus on making use of audio as a complementary source of information for action anticipation.

3. Methodology

Our architecture which is displayed in Figure 2 consists of three exchangeable components: Modality specific feature extractors $f_{\Pi}^{m_j}, j \in \{1, \dots, M\}$, a cross-modal fusion module g_{Φ} and an anticipation module a_{Ω}. Since this work analyzes multi-modal fusion on frozen features, we assume all f_{Π} to have pretrained frozen weights and therefore refer to Section 4.2 for more details on the specific feature sets used for our experiments. Our proposed fusion modules are presented in Section 3.2. We follow [20] and use a variation of the GPT-2 [40] model as feature anticipation module to predict $\hat{z}_{i+1} = a_{\Omega}(z_i), i \in \{1, \dots, T\}$.

3.1. Problem statement

In this work, we follow the anticipation setup defined in [8, 9]. As illustrated in Figure 1, the action anticipation task aims to predict an action starting at time τ_s by observing a video segment of length τ_o. The observation segment is τ_a seconds preceding the action, i.e., from time $\tau_s - (\tau_a + \tau_o)$ to $\tau_s - \tau_a$, where τ_a denotes the "anticipation time", i.e., how many seconds in advance actions are to be anticipated. The anticipation time τ_a is usually fixed for each dataset, whereas the length of the observation segment is typically dependent on the individual method. In

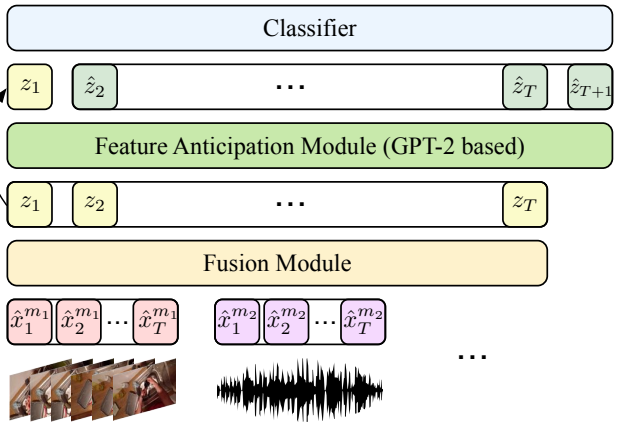

Figure 2: Architecture of AFFT. The feature encoders are omitted, we directly list the feature vectors $\hat{x}^M$. A fusion module combines the modality specific feature vectors. The feature anticipation module then predicts the features of the next time step, followed by a linear classifier.

our experiments we assume T temporally sequential input observations $x_i^{m_j}, i \in \{1, \dots, T\}, j \in \{1, \dots, M\}$ which describe the observation time τ_o for each of the M available modalities. The anticipated action is defined to be at time step $T+1$ without observation and label y_{T+1}. Depending on the dataset, the preceding observations might additionally be labelled with y_i. Since this work is aimed at feature based modality fusion, we assume fixed feature extractors f_Π and define the individual extracted features as $\hat{x}_i^{m_j} = f_\Pi^{m_j}(x_i^{m_j})$ and the collection of all $T \times M$ features for an input sample as $\hat{x}^M$.

3.2. Cross-modal fusion

Time-decoupled feature fusion In order to fuse the features $\hat{x}_i^M$ on each individual time-step separately, we apply L consecutive transformer encoder blocks as used in [10] with dimensionality d and k attention heads, this module is displayed on the left of Figure 3. We found that modality-wise positional embeddings do not yield an improvement of performance, presumably since the modality specific features are already easily separable in feature space. We do ablate the usage of a modality agnostic learnable token $\hat{x}^\Lambda$, similar to the concept of a learnable class-token used in [10]. The module with the prepended learnable token x^Λ is referred to as *Self-Attention Fuser (SA-Fuser)*. Without this learnable token we average the resulting output tokens $z_i^{m_j}$. We consider the usage of the learnable token as default, experiments without token are marked as such.

Temporal feature fusion The *Temporal Self-Attention Fuser (T-SA-Fuser)* which is displayed in the middle of Figure 3 follows the paradigm of the *SA-Fuser*, but instead of fusing multi-modal features per time step, all modality features for all time steps are used to provide all output features $z = g(\hat{x}^M)$ at once. A learnable positional embedding p_i is used to encode the temporal positions for each modality and an attention mask enforces that an output feature at temporal position i only attends to previous or concurrent multi-modal features. Instead of a single modality agnostic token, we provide a learned token for each time step provided to the module. Learned positional embeddings are added to the tokens of each time step to allow the model to differentiate them.

Temporal cross-attention feature fusion Our third fusion module, which is displayed on the right of Figure 3, is inspired by [26] and follows a different paradigm. Instead of providing all modalities at once, we iteratively enrich a main modality (RGB in our experiments) with information from other modalities. Instead of L transformer encoder blocks, $(M-1)$ transformer decoder blocks [46] are used. Following the decoder architecture, the RGB features $\hat{x}^{RGB}$ are provided as main input which provides the queries for the multi-head cross-attention and each block makes use of another modality $\hat{x}^{m_j}$ as second decoder input which provides the keys and values. Positional embeddings are added to all modality features. We do not use additional tokens, but rather directly predict the fused features z. We refer to this module as *Cross-Attention Fuser (CA-Fuser)*

3.3. Feature anticipation and classification

After different modality features get fused by the fusion module, a variation of the GPT-2 [40] model is used to predict the future features $\hat{z}_{i+1} = a_\Omega(z_i), i \in \{1, \dots, T\}$, following [20]. To encode the temporal ordering and obtain generative ability, learnable positional embeddings and a temporal attention mask are used. Based on the anticipated features $\hat{z}$ we define a classification head h, a single linear layer followed by a softmax activation function. The anticipation result is based on the predicted future feature, so $\hat{y}_i = h(\hat{z}_i)$ and the final anticipation result $\hat{y}_{T+1} = h(\hat{z}_{T+1})$.

3.4. Loss functions

Our loss functions follow the setting of [20]. We apply three losses $\mathcal{L} = \mathcal{L}_{\text{next}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{feat}}$. $\mathcal{L}_{\text{next}}$ is defined on $\hat{y}_{T+1}$ and y_{T+1} according to the task of action anticipation. Since the network output does not only provide features $\hat{z}_{T+1}$ for the anticipated next action but also for the preceding time steps $i \in \{1, \dots, T\}$, $\mathcal{L}_{\text{cls}}$ evaluates the action classification performance of these preceding features, so on $\hat{y}_i = h(\hat{z}_i)$ and y_i. Both are cross-entropy losses. $\mathcal{L}_{\text{feat}}$ is the mean squared error between predicted and fused features $\hat{z}_i$ and z_i.

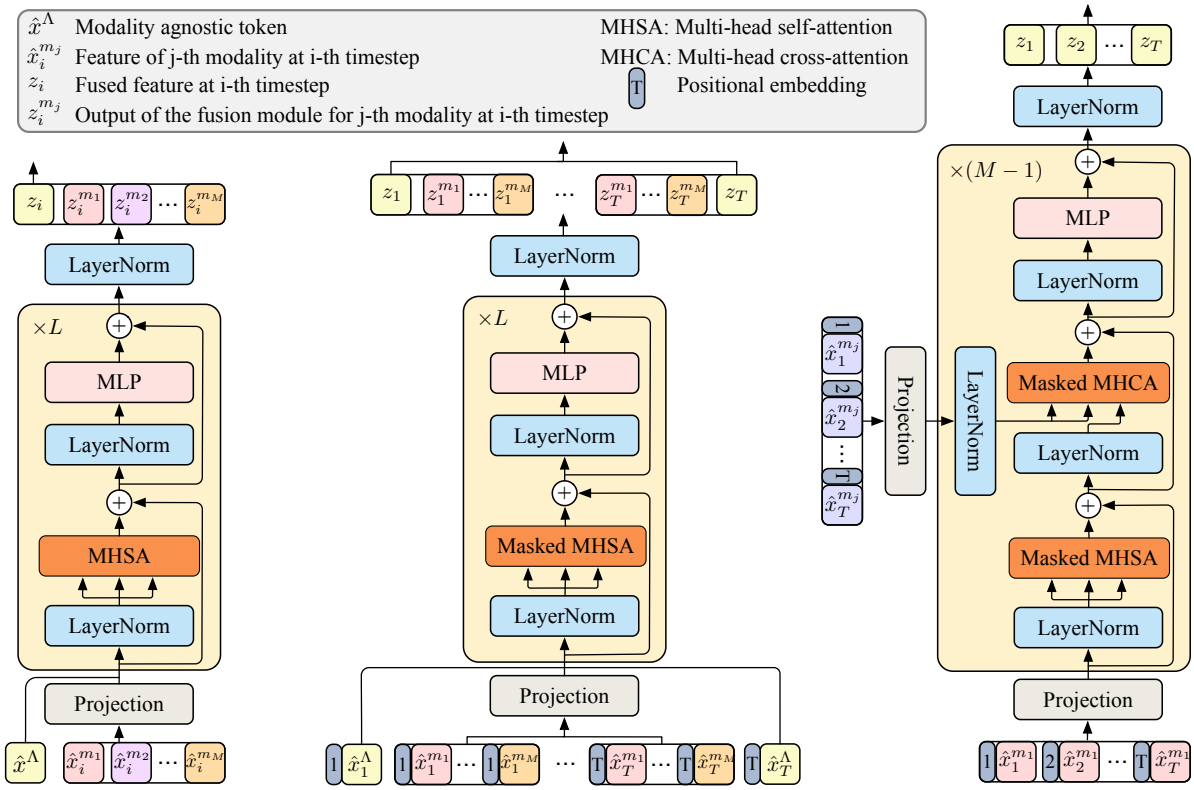

Figure 3: The *SA-Fuser* on the left applies Transformer Encoder blocks at individual time steps while the *T-SA-Fuser* in the middle and the transformer decoder based *CA-Fuser* on the right perform fusion on the whole temporal sequence at once.

4. Experimental Setup

In order to investigate the influence of the different fusion strategies and evaluate the proposed method for the action anticipation task, we train and evaluate our methods on two different datasets (discussed in detail in Section 4.1). To allow a fair comparison with prior work, we first use pre-extracted TSN features [47] as input features for both datasets provided by [16]. To investigate the impact of the audio modality for action anticipation, we train a TSN audio action recognition model following [28] and extract its features for fusion with other modalities. In order to show the generalization of our proposed fusion method, we extract alternative RGB features from a recent state-of-the-art visual model, OMNIVORE [22]. Information regarding feature extraction is discussed in detail in Section 4.2. All experiments follow the training procedure described in Section 4.3.

4.1. Datasets and metrics

We perform experiments on two large-scale egocentric (first-person) video datasets: EpicKitchens-100 [9] and EGTEA Gaze+ [32]. EpicKitchens-100 consists of 700 long unscripted videos of cooking activities totalling 100 hours. It contains 90.0K action annotations, 97 verbs, and 300 nouns. We considered all unique (*verb*, *noun*) pairs in the public training set, obtaining 3,807 unique actions. We use the official train, val and test splits to report performance. The test evaluation is performed on a held-out set through a submission to the official challenge server. EGTEA Gaze+ is another popular egocentric action anticipation dataset. It contains 10.3K action annotations, 19 verbs, 51 nouns and 106 unique actions.

We report class mean top-5 recall [14] for EpicKitchens-100, a class-aware metric in which performance indicators obtained for each class are averaged to obtain the final score, accounting for the multi-modality in future predictions and class imbalance in a long-tail distribution. For EGTEA Gaze+, we report top-1/5 and class mean top-1. As some prior works report their results averaged across the official three splits, and some evaluate their methods on the first split only, we test our method using both recipes.

4.2. Uni-modal features

RGB. We compare two types of RGB features, the commonly used TSN features [47] provided by [16] and Swin transformer [34] features which we extracted with OMNI-

VORE [22] to represent more recent transformer based approaches. Both feature extractors are trained for action recognition. While TSN features are extracted by applying TSN on each frame, we extract Swin features by feeding 32 consecutive past frames totalling 1.067s video with a frame rate of 30fps to the OMNIVORE model for each timestamp.

Audio. Following [28], We extract 1.28s of audio, convert it to single-channel, and resample it to 24kHz. We then convert it to a log-spectrogram representation using an STFT of window length 10ms, hop length 5ms and 256 frequency bands, resulting in a 2D spectrogram matrix of size 256×256, after which we compute the logarithm. Different from [28], we extract audio in an online manner, i.e., we extract the past audio segment for each timestamp, prohibiting the model to have access to the future, which is the prerequisite for the anticipation task. We feed such matrices to the TSN network, train it for the action recognition task and extract features for our work.

Objects and optical flow. We use the existing object and optical flow features provided by [16]. Object representations are obtained by accumulating the confidence scores of all bounding boxes predicted by a Faster R-CNN [41] for each object class. Optical flow features are extracted by feeding 5 consecutive past frames of horizontal and vertical flow, forming a tensor with 10 channels, to a TSN model trained for action recognition.

4.3. Implementation details

Architecture details. For our AFFT model we use the marked default hyper parameters from Table 2. For EGTEA Gaze+, we reduce the number of layers of the fuser and the future predictor to 2, since EGTEA Gaze+ is relatively small compared to EpicKitchens-100. We employ a linear projection layer for modality features that are not in alignment with the hidden size of the fuser. To match the hidden dimension used in the future predictor, another linear layer is employed to project the fused modality features.

Training & testing. We sample all modality features at 1 fps, resulting in a sequence of feature vectors whose length corresponds to observation time τ_o. Default observation time is 10s, the other observation lengths are analyzed in Section 5.3. We train our models with SGD+momentum using 10^{-6} weight decay and 10^{-3} learning rate for 50 epochs, with 20 epochs warmup [24] and 30 epochs of cosine annealed decay, following [20]. We use mixup data augmentation [51] with $\alpha = 0.1$. Default settings for dropout and the stochastic depth regularization technique [25] are listed in Table 2. Following standard practice [16, 20, 48], our model is optimized to predict the action label during training and marginalize the output probabilities to obtain the verb and noun predictions in testing.

	Fusion strategy	Act.
Score	Average	16.4
	Weighted average	17.3
	MATT	12.2
Feature	SA-Fuser (w/o Token)	17.1
	SA-Fuser	**18.0**
	T-SA-Fuser	15.2
	CA-Fuser	16.6

Table 1: Comparison of fusion strategies. The results are based on all modalities with RGB-Swin features. We refer to the model using the *SA-Fuser* as our method AFFT.

5. Results

In Section 5.1 we ablate the proposed fusion architectures. Continuing with the best architecture, we find optimized hyper parameters in Section 5.2 and the optimal temporal context in Section 5.3. In Section 5.4 we analyze the contribution of individual modalities to the final model performance and in Section 5.5 our models are compared against state-of-the-art feature based action anticipation models on EpicKitchens-100 and EGTEA Gaze+. The models trained with RGB-TSN and RGB-Swin features are referred to AFFT-TSN and AFFT-Swin respectively.

5.1. Fusion strategies

We evaluate the fusion architectures presented in Section 3.2 against score fusion based methods and evaluate which of our strategies proves best for multi-modal fusion. Table 1 lists all methods. In our comparison we include *Modality Attention (MATT)* [16], a learned score fusion weighting method, but find it to be lacking in our setting. For score averaging and weighted averaging, we choose the same setting as [20], verifying their results. Combining temporal and modality attention as done with *T-SA-Fuser* performs worst in our feature fusion models, which we assume to be caused by the complexity of this process. *CA-Fuser* introduces an inductive bias by introducing a new modality with each consecutive block, splitting the process of attention into separate smaller problems instead of presenting all temporal and modality tokens at once. Our best approach *SA-Fuser* on the other hand is even simpler, since it splits the problem along time-steps and only attends over the modality tokens. Temporal attention is then performed in a completely separate step with the GPT-2 based future predictor. We believe this reduced complexity to be the mechanism which leads to optimal performance of our final model. For further experiments we use the *SA-Fuser* as our default fusion module.

Projection	Act.
Lin.	17.6
Lin. (sparse)	**18.0**
Lin. + ReLU	17.1
GLU [36]	17.8

(a) Projection layer.

Dim.	Act.
512	16.7
768	17.2
1024	**18.0**
1280	18.0
2048	16.9

(b) Dimension.

Heads	Act.
4	18.0
8	**18.4**
16	17.4

(c) No. of heads.

Layers	Act.
2	17.9
4	17.3
6	**18.0**

(d) No. of layers.

Regularization	Act.
no regularization	15.8
stochastic depth (0.1)	16.2
drop. & stoch. depth (0.1)	**18.0**

(e) Regularization.

Table 2: Fuser architecture ablation on the validation set of EpicKitchens-100. Default settings are marked in gray.

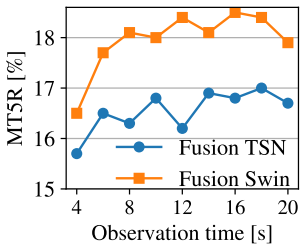

Figure 4: Impact of temporal context on the validation set of EpicKitchens-100. Our method leverages long-term dependencies to improve anticipation performance.

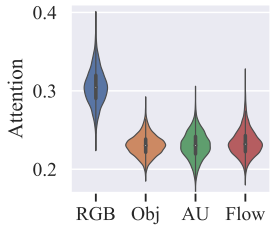

Figure 5: Modality attentions of AFFT-Swin on the validation set of EK-100. Our method learns to pay more attention to RGB without any supervision.

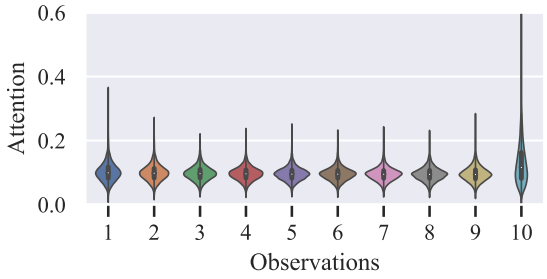

Figure 6: Temporal attentions of AFFT-Swin over all samples of the validation set of EK-100. Our method attends not only to the recent past, but also to the entire past frames.

5.2. Architecture ablations

In Table 2, we ablate different hyper parameters of our architecture. The default parameters are marked with grey table cells, the best values are typed in boldface.

Projection layer and common dimensionality. The dimension of all multi-modal input features must coincide. This could be achieved using a simple linear layer, a linear layer with ReLU activation function [27, 23] or a gated linear projection [36, 44], listed in Table 2a. We add an additional variant *sparse linear*, meaning a linear layer is only applied for features which have a different dimension than the desired common dimension and show that it outperforms other projection methods. In Table 2b, we examine how the projection dimension influences performance. We find a dimensionality of 1024 to be optimal, a higher dimension presumably decreases performance due to the increased number of parameters and overfitting effects.

Attention heads and encoder blocks. We compare the impact of different head numbers of the encoder multi-head attention in Table 2c, the number of encoder blocks is analyzed in Table 2d, we find eight heads and six consecutive encoder blocks to be best.

Effect of regularization. We ablate using either no dropout and no stochastic depth [25] (i.e. no regularization) or using stochastic depth with maximal layer dropping probability of 0.1. Results in Table 2e show that both dropout and stochastic depth regularization are very beneficial.

5.3. Impact of temporal context.

To study the ability of modeling sequences of long-range temporal interactions, we train and test the model with different lengths of temporal context, i.e., observation time τ_o. As seen in Figure 4, as more frames of context are incorporated, the performance improves for both, AFFT-TSN and AFFT-Swin. The gains are especially pronounced when trained using RGB-Swin features ($16.5 \rightarrow 18.5 = 2.0 \uparrow$) vs. RGB-TSN features ($15.7 \rightarrow 17.0 = 1.3 \uparrow$). To further explore how the temporal context is utilized, following [20], we extract temporal attentions from the last layer of the feature anticipation module for all samples in the validation set of EpicKitchens-100, average them over heads and visualize them in Figure 6. The anticipation module learns to attend to visual features in the recent past, showing that the nearest past frames provide crucial keys for predicting future actions. This aligns with previous work [29, 42] which reflects the importance of the recent past in designing anticipation models. However, while the median attention values of more distant past frames are smaller (close to 0.1), the attention distribution is significantly scattered, indicating that the model can choose to attend to important actions not only from the recent past, but also from the entire observation time, as illustrated in an example in Figure 8. Here the model attends to an early time step in the middle of the observation which shows the opening of a fridge in order to predict the the future action 'close fridge'. Results for AFFT-TSN are listed in the supplementary.

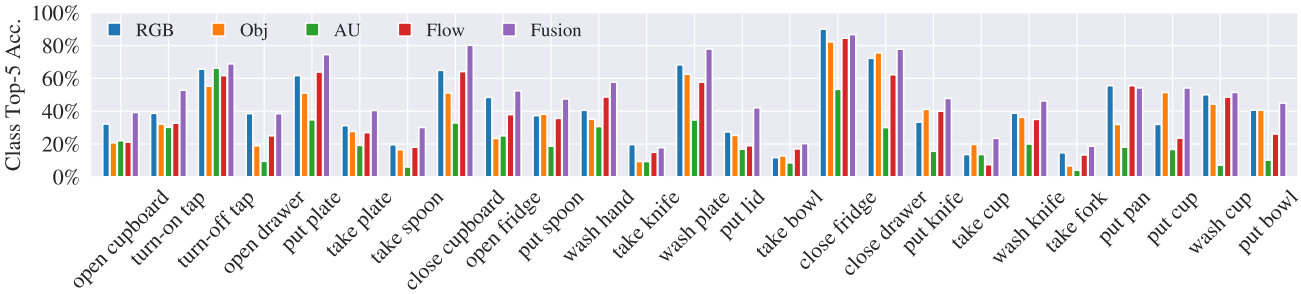

Figure 7: Per-class top-5 accuracy of fusion (AFFT-TSN) and single modalities for the largest 25 actions in the validation set of EpicKitchens-100. The classes are presented in the order of number of samples per class, from left to right. For most classes the fusion method provides significantly better results over the single modalities.

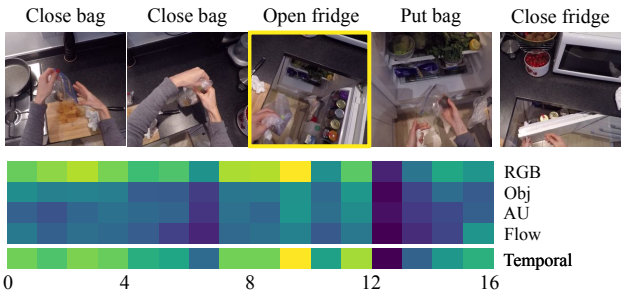

Figure 8: Qualitative results on EpicKitchens-100. The horizontal and vertical axes indicate the index of past frames as well as the modality. The closer the color is to yellow, the higher the attention score. A video frame is highlighted with a yellow box when the attention score of the frame is highly activated.

Mod.	Backbone	Act.
RGB	TSN	13.2
RGB	Swin	16.1
Obj	F. R-CNN	9.9
AU	TSN	5.3
Flow	TSN	7.5

(a) Results of individual modalities.

RGB	TSN	Swin
Other	Act.	Act.
Obj	15.9	16.7
AU	15.4	16.8
Flow	15.2	16.5
Obj+Flow	16.2	17.6
Obj+AU+Flow	16.8	18.0

(b) Results of multiple modalities combined with RGB.

Table 3: Impact of individual modalities on the validation set of EpicKitchens-100. Compared to other modalities, RGB performs significantly better, particularly on features extracted by Swin. The proposed fusion method benefits from multi-modal inputs. The more modalities are provided, the better the anticipation model performs.

5.4. Modality contributions

As shown in Table 3a, visual modalities, especially RGB, have higher performances than audio, also observable in Figure 7. Benefiting from a larger model capacity and better representative ability, RGB features extracted with an Omnivore-pre-trained Swin-Transformer perform significantly better than TSN features. Results in Table 3b show that the anticipation performance keeps increasing when additional modalities are introduced for both kinds of RGB features. In particular, AFFT-TSN and AFFT-Swin have gains of 3.6% and 1.9% over their uni-modal RGB performances in Table 3a, respectively. Per-class top-5 accuracies, for individual modalities as well as for our fusion model (AFFT-TSN) trained on all four modalities, can be seen in Figure 7. The fusion model outperforms uni-modal models for most classes, often by a significant margin. Results of AFFT-Swin are shown in the supplementary material. To analyze the contribution of our extracted audio features, we conduct experiments with the visual modalities (RGB+Obj+Flow) only and compare them with models trained on all four modalities, which results in an increase of 0.6% (AFFT-TSN) and 0.4% (AFFT-Swin) in mean top-5 action anticipation accuracy as seen in Table 3b. To further validate the benefit of audio, we compute a confusion matrix with the utilization of audio for the largest-15 action classes, following [28], which we list in the supplementary. To better understand how the fusion module models relative importance of different modalities, we visualize the learned modality attentions of AFFT-Swin in Figure 5. Specifically, we use attention rollout [1] to aggregate attention over heads and layers. As shown in the figure, RGB has gained the most attention, indicating the modality which contributes the most for the anticipation task (as seen in Table 3a) will be automatically utilized most by the fusion module, as would be expected. Figure 5 also shows that the attention distributions of all modalities spread widely, showing that the model learns to adjust the relative importance of individual modalities based on each sample.

	Method	Overall			Unseen Kitchen			Tail Classes		
		Verb	Noun	Act.	Verb	Noun	Act.	Verb	Noun	Act.
Val	chance	6.4	2.0	0.2	14.4	2.9	0.5	1.6	0.2	0.1
	AVT+ [20]	28.2	32.0	15.9	29.5	23.9	11.9	21.1	25.8	14.1
	MeMViT [48]	32.2	37.0	17.7	28.6	27.4	15.2	25.3	31.0	15.5
	RULSTM [16]	**27.8**	30.8	14.0	**28.8**	**27.2**	14.2	**19.8**	22.0	11.1
	TempAgg [42]	23.2	31.4	14.7	28.0	26.2	14.5	14.5	22.5	11.8
	AVT+-TSN [20]	25.5	31.8	14.8	25.5	23.6	11.5	18.5	25.8	12.6
	Ours-TSN	21.3	32.7	16.4	24.1	25.5	13.6	13.2	25.8	14.3
	Ours-TSN$^+$	22.3	31.5	17.0	23.8	25.3	14.0	14.6	23.6	15.0
	Ours-Swin	23.4	33.7	17.6	24.5	25.4	15.2	15.6	26.5	15.3
	Ours-Swin$^+$	22.8	**34.6**	**18.5**	24.8	26.4	**15.5**	15.0	**27.7**	**16.2**
Test	chance	6.2	2.3	0.1	8.1	3.3	0.3	1.9	0.7	0.0
	AVT+ [20]	25.6	28.8	12.6	20.9	22.3	8.8	19.0	22.0	10.1
	RULSTM [16]	**25.3**	26.7	11.2	19.4	26.9	9.7	**17.6**	16.0	7.9
	TempAgg [42]	21.8	30.6	12.6	17.9	27.0	10.5	13.6	20.6	8.9
	TCN-TSN [50]	20.4	26.6	10.9	17.9	26.9	11.1	11.7	15.2	7.0
	TCN-TBN [50]	21.5	26.8	11.0	**20.8**	**28.3**	**12.2**	13.2	15.4	7.2
	Ours-TSN$^+$	19.4	28.3	13.4	14.0	24.2	9.9	12.0	19.5	10.9
	Ours-Swin$^+$	20.7	**31.8**	**14.9**	16.2	27.7	12.1	13.4	**23.8**	**11.8**

Table 4: Comparison of state-of-the-art methods on the validation and test set of EpicKitchens-100. Our models set a new state of the art. The numbers in bold-face indicate the highest score. All methods use all modalities provided by [16], except for MeMViT which uses RGB only. TempAgg and the ones marked with $^+$ additionally use interacting hand-object bounding boxes and audio, respectively.

5.5. Comparison to the state-of-the-art

Our final models follow the default hyper parameters from Table 2. On EpicKitchens-100, AFFT-TSN and AFFT-Swin use observations of 18s and 16s respectively, while the default observation time (10s) is used for EGTEA Gaze+. For the comparisons, we distinguish between training with frozen backbones (i.e., training on frozen features) and training with fine-tuned backbones (marked with gray font). In all tables in this section, the main metrics used to rank methods for these datasets are highlighted.

In Table 4, we compare our method with state-of-the-art methods on EpicKitchens-100. The table is divided into two compartments according to the validation and test splits. On the validation split, our AFFT-TSN outperforms other fusion methods with a large margin ($14.8 \rightarrow 16.4 = 1.6 \uparrow$) with the exact same features provided by [16]. With the addition of audio, the performance is further improved by 0.6%. AFFT-Swin$^+$ which uses Omnivore features outperforms the current state-of-the-art model MeMViT by 0.8% mean top-5 ratio action anticipation performance on the val split without the need to fine-tune the backbone network. Consistent with the results on validation split, our method also outperforms prior fusion methods on the test set of EpicKitchens-100. As shown in bottom compartment in Table 4, we get the largest gains on tail classes, for which our method proves particularly effective. Note that Table 4 lists peer-reviewed results, only. In our supplementary we also list results of the EpicKitchens-Challenge, which holds many non-peer-reviewed results, often created with model ensembling of various methods.

Method	Top-1			Class mean @1			Top-5
	Verb	Noun	Act.	Verb	Noun	Act.	Act.
I3D-Res50 [7]	48.0	42.1	34.8	31.3	30.0	23.2	-
FHOI [33]	49.0	45.5	36.6	32.5	32.7	25.3	-
AVT [20]	54.9	52.2	43.0	49.9	48.3	35.2	-
RULSTM [16]	-	-	-	-	-	-	71.84*
ImagineRNN [49]	-	-	-	-	-	-	72.32*
AVT (TSN) [20]	51.7	50.3	39.8	41.2	41.4	28.3	-
AFFT-TSN (Ours)	**53.4**	**50.4**	**42.5**	**42.4**	**44.5**	**35.2**	**72.47***

Table 5: Comparison to the state-of-the-art methods on EGTEA Gaze+ with $\tau_a = 0.5s$. Results marked with * are averaged across the three official splits, while others are based on split 1 only. We use the same input modalities as RULSTM. More details on the used modalities of each method can be found in the supplementary material.

Next we evaluate our method on EGTEA Gaze+, shown in Table 5. Following prior works [33, 20], we set the anticipation time τ_a to 0.5s. As some prior works report the results averaged across the three official splits, while others test on split 1 only, we evaluate our methods using both recipes. Using fixed features, AFFT-TSN outperforms prior works using both recipes, especially for class mean top-1.

6. Conclusion and Future Work

This work presents Anticipative Feature Fusion Transformer (AFFT), an attention based multi-modal feature fusion method for action anticipation. Extensive ablations demonstrate the improved performance of our approach compared to basic score fusion or other multi-modal fusion methods and in state-of-the-art comparisons AFFT outperforms existing approaches on EpicKitchens-100 and EGTEA Gaze+. Our method can easily be combined with various feature extractors and is extensible to new modalities without architectural changes. Given this extensibility we hope to provide a framework for multi-modal action anticipation for other researchers and aim to experiment on additional modalities like body poses and object hand interactions ourselves, in the future.

Acknowledgements This work was supported by the JuBot project which was made possible by funding from the Carl-Zeiss-Foundation. This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, 2018.

[3] Relja Arandjelović and Andrew Zisserman. Look, Listen and Learn. In *ICCV*, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, pages 435–451, 2018.

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, volume 29, 2016.

[6] Floyd E Bloom, Arlyne Lazerson, Laura Hofstadter, et al. *Brain, mind, and behavior*, volume 300. Freeman New York, 1988.

[7] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, number arXiv:1705.07750, 2017.

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

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *TPAMI*, 43(11):4125–4141, 2020.

[10] 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 *ICLR*, 2021.

[11] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *SIGGRAPH*, 2018.

[12] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - Anticipating Temporal Occurrences of Activities. In *CVPR*, 2018.

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[14] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging Uncertainty to Rethink Loss Functions and Evaluation Measures for Egocentric Action Anticipation. In *ECCVW*, 2018.

[15] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017.

[16] Antonino Furnari and Giovanni Farinella. What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention. In *ICCV*, 2019.

[17] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. RED: Reinforced Encoder-Decoder Networks for Action Anticipation. In *BMVC*, 2017.

[18] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *CVPR*, pages 324–333, 2019.

[19] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020.

[20] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021.

[21] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, pages 971–980, 2017.

[22] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022.

[23] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future Transformer for Long-term Action Anticipation. In *CVPR*, 2022.

[24] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[25] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016.

[26] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP*, pages 3507–3511, 2020.

[27] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a Little Help from my Temporal Context: Multimodal Egocentric Action Recognition. In *BMVC*, 2021.

[28] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. In *ICCV*, 2019.

[29] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-Conditioned Action Anticipation in One Shot. In *CVPR*, June 2019.

[30] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization. In *NeurIPS*, 2018.

[31] Bruno Korbar, Du Tran, and Lorenzo Torresani. SCSampler: Sampling Salient Clips From Video for Efficient Action Recognition. In *ICCV*, 2019.

[32] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, pages 619–635, 2018.

[33] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting Human-Object Interaction: Joint Prediction of Motor Attention and Actions in First Person Video. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020.

[34] 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 *CVPR*, pages 10012–10022, 2021.

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022.

[36] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with Context Gating for video classification. In *CVPR Workshop*, 2017.

[37] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention Bottlenecks for Multimodal Fusion. In *NeurIPS*, 2021.

[38] Andrew Owens and Alexei A. Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018.

[39] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, pages 2405–2413, 2016.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, volume 28, 2015.

[42] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal Aggregate Representations for Long-Range Video Understanding. In *ECCV*, 2020.

[43] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *TPAMI*, 43(5):1605–1619, 2019.

[44] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at Once - Multi-Modal Fusion Transformer for Video Retrieval. In *CVPR*, 2022.

[45] Durk Talsma. Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, 9:19, 2015.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017.

[47] 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*, 2016.

[48] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition. In *CVPR*, 2022.

[49] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to Anticipate Egocentric Actions by Imagination. *TIP*, 2021.

[50] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. Multi-Modal Temporal Convolutional Network for Anticipating Actions in Egocentric Videos. In *CVPR Workshop*, 2021.

[51] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[52] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, pages 3550–3558, 2018.