Supplementary Material

In this supplementary, we provide additional experiments and evaluations which did not fit in the main paper.

A. Temporal and modality contributions

In Section 5.3, Figure 6 displays temporal attention values for AFFT-Swin. In Figure S10 we show the same evaluation for the extracted TSN features.

Likewise, Figure 5 displays distributions of attention values over modalities using RGB-Swin features which is completed by Figure S9 which displays the the same results for RGB-TSN features.

Figure S13 shows per-class top-5 accuracy for the 30 action classes with most samples in the EpicKitchens-100 validation set, based on RGB-Swin Features. A similar chart is displayed in Figure 7 for TSN features. Note, that for such high frequent classes, performance is significantly higher than in the overall dataset. Still, our method does not only perform well for high frequent classes, but also shows significantly improved results for tail classes, as can be seen in Table 4.

B. Confusion Matrix

We follow the work of Kazakos et al. [28] and evaluate the contribution of the audio modality, specifically. Figure S11 shows the confusion matrices for the 15 most frequent action classes on the left and displays the difference to the confusion matrix without the audio modality on the right. While this Figure reflects the limited contributions of audio which are also visible in Table 3b, especially for Swin-Features, an increase of performance on the diagonal can be noted.

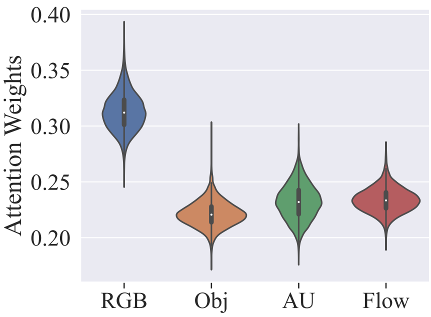


Figure S9: Modality attentions of AFFT-TSN on the validation set of EpicKitchens-100. This figure is the counterpart to Figure 5, which describes the same evaluation on Swin RGB features.

C. Qualitative Results

We plot additional visualization results of modality and temporal attentions in Figure S12. The model used to generate such plots corresponds to the AFFT-Swin in Table 4. Each subfigure contains sampled frames showing temporal action evolution and modality and temporal attention map visualizations below. The frame receiving the most temporal attention is highlighted with a yellow box. From this experiment, we find that the proposed method attends dynamically to the multi-modalities and different past time steps to predict the future action, which demonstrates that our method successfully leverages long-term dependencies using multi-modal information for key frame detection and action anticipation.

Method	Overall			Unseen Kitchen			Tail Classes		
	Verb	Noun	Act.	Verb	Noun	Act.	Verb	Noun	Act.
AVT++ [20]	26.7	32.3	16.7	21.0	27.6	12.9	19.3	24.0	13.8
allenxuuu	29.9	30.4	17.4	25.1	26.1	14.1	24.6	23.7	14.3
PCO-PSNRD	30.9	41.3	18.7	25.7	35.4	16.3	25.0	35.4	16.1
ICL-SJTU	**42.0**	35.7	19.5	**33.4**	26.8	15.9	**41.0**	33.2	16.9
NVIDIA-UNIBZ	29.7	38.5	19.6	23.5	35.2	16.4	23.5	31.1	16.6
SCUT	37.9	**41.7**	**20.4**	27.9	**37.1**	**18.3**	32.4	**36.1**	**17.1**

Table S6: Current leaders in the EpicKitchens-100 action anticipation challenge. The numbers in bold-face indicate the highest score.

D. EpicKitchens-100 challenge

Table S6 lists results from the EpicKitchens-100 action anticipation challenge. This table relates to the test results in Table 4. Entries to the challenge typically significantly surpass single-method performances, since it is common to ensemble differently trained models or results from different methods. We list this table separately, since its ensem-

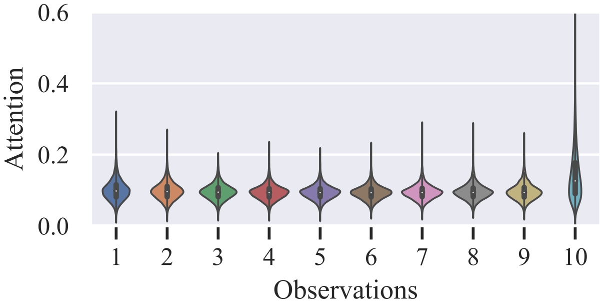


Figure S10: Temporal attentions of AFFT-TSN over all samples of the validation set of EpicKitchens-100. This figure describes a similar pattern and validates the evaluation on Swin features in figure 6.

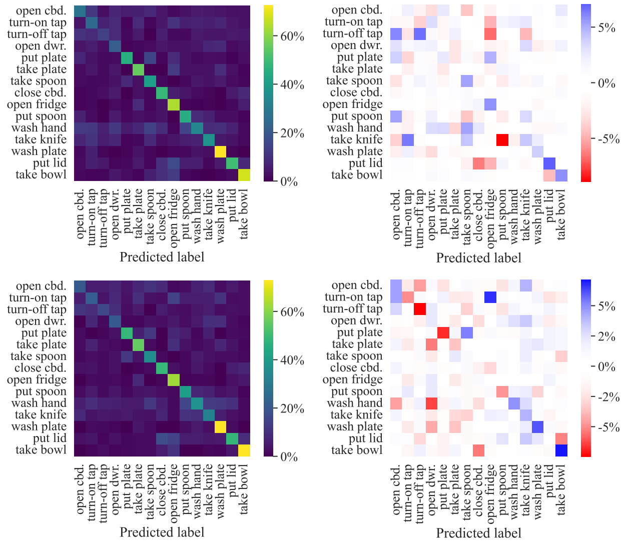
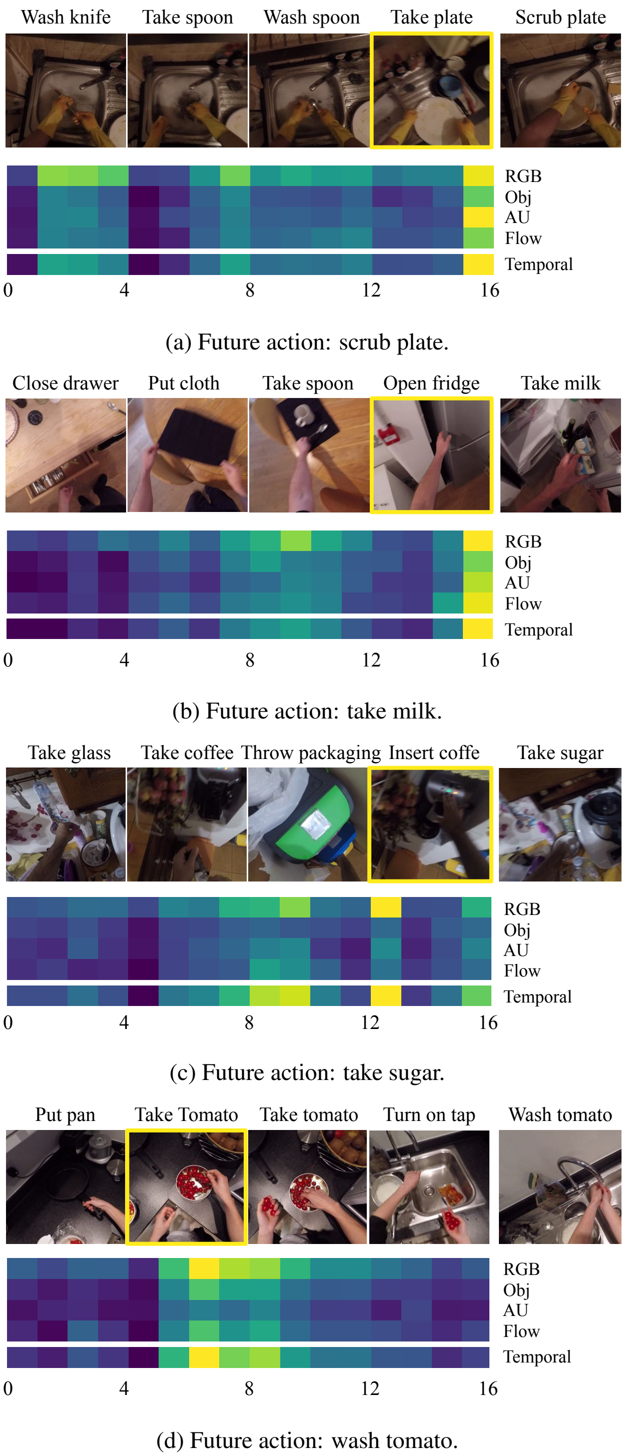


Figure S11: Confusion matrix for the largest-15 action classes in the validation set of EpicKitchens-100, with audio (left), as well as the difference to the confusion matrix without audio (right). From top to bottom, results of AFFT-TSN and AFFT-Swin are shown. An increase (blue) in confidence along the diagonal, especially obvious in the upper right figure, demonstrates the benefit of audio modality for egocentric action anticipation.

bled results can not be directly compared and did not undergo peer review.

E. Details of used modalities on EGTEA Gaze+

While a single RGB modality is used for I3D-Res50, FHOI [33] adopts intentional hand movement as a feature representation, and jointly models and predicts the egocentric hand motion, interaction hotspots and future action. On the other hand, RULSTM [16] and ImagineRNN [49] make use of multiple modalities, i.e., RGB and optical flow, to further improve the anticipation performance of the next action. We note that the modalities used in AVT are ambiguous. We follow RULSTM and ImagineRNN and use RGB and optical flow as the input modalities for our method.



(a) Future action: scrub plate.

(b) Future action: take milk.

(c) Future action: take sugar.

(d) Future action: wash tomato.

Figure S12: Qualitative results on EpicKitchens-100. The horizontal and vertical axes indicate the index of the past frames and the modality as well as temporal attention scores, respectively. The closer the color is to yellow, the higher the attention score. We highlight a video frame with a yellow box when the attention score of the frame is highly activated.

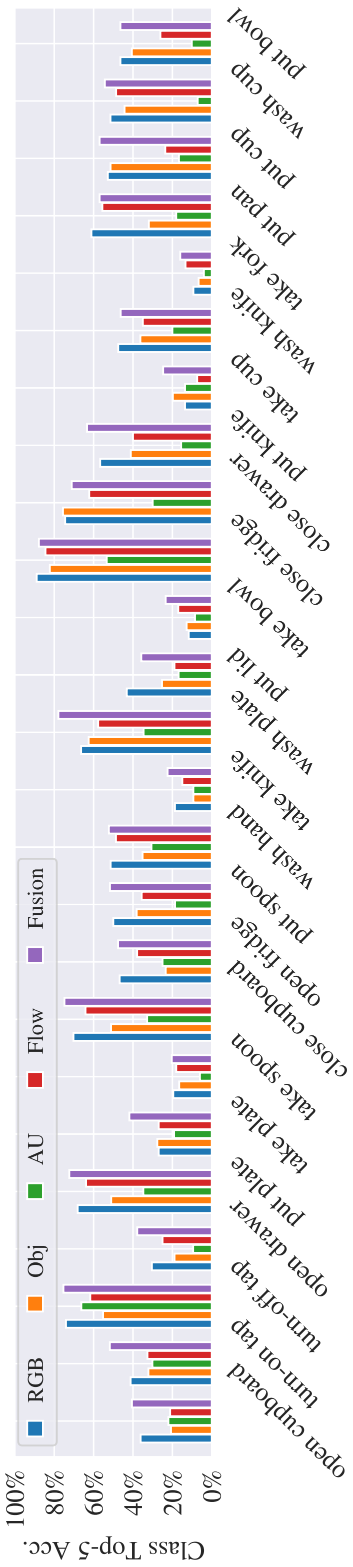


Figure S13: Per-class top-5 accuracy of fusion (AFFT-Swin) and single modalities for the largest-25 actions in the validation set of EpicKitchens-100. The classes are presented in the order of sample frequency, from left to right. For most classes, the fusion method provides superior results to the single modalities.