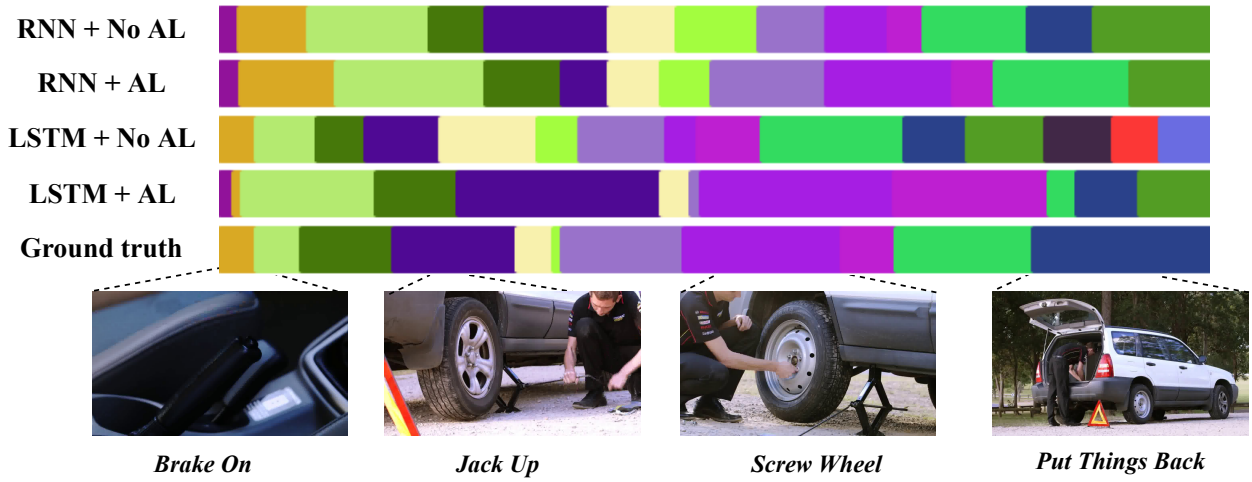


Figure 4: **Ablative Studies**: Illustrative comparison of variations of our approach, using RNNs and LSTMs with and without adaptive learning on the INRIA Instructional Videos Dataset on a video with ground truth *Change Tire*. It can be seen that complex visual scenes with activities of shorter duration pose a significant challenge to the proposed framework and cause fragmentation and over segmentation. However, the use of adaptive learning helps alleviate this to some extent. Note: Temporal segmentation time lines are shown without the background class for better visualization.

| Supervision | Approach | MoF | IoU |
|---|---|---|---|
| Full | SVM [19] | 15.8 | - |
| | HTK(64)[20] | 56.3 | - |
| | ED-TCN[27] | 43.3 | 42.0 |
| | TCFPN[10] | 52.0 | **54.9** |
| | GRU[29] | **60.6** | - |
| Weak | OCDC[6] | 8.9 | 23.4 |
| | ECTC[16] | 27.7 | - |
| | Fine2Coarse[28] | 33.3 | **47.3** |
| | TCFPN + ISBA[10] | **38.4** | 40.6 |
| None | KNN+GMM[30] | 34.6 | **47.1** |
| | Ours (LSTM + AL) | **42.9** | 46.9 |

Table 1: Segmentation Results on the Breakfast Action dataset. MoF refers to the Mean over Frames metric and IoU is the Intersection over Union metric.

proach [30], requires the number of clusters (from ground truth) to achieve the performance whereas our approach does not require such knowledge and is done in a streaming fashion. Additionally, the weakly supervised methods [16, 28, 10] require both the number of actions as well as an ordered list of sub-activities as input. ECTC [16] is based on discriminative clustering, while OCDC [6] and Fine2Coarse [28] are also RNN-based methods.

**50 Salads Dataset** We also evaluate our approach on the 50 Salads dataset, using only the visual features as input. We report the Mean of Frames (MoF) metric for fair comparison. As can be seen from Table 2, the proposed approach significantly outperforms the other unsupervised approach, improving by over 11%. We also show the performance of the frame-based classification approaches VGG and IDT [21] to show the impact of temporal modeling. It should be noted that the fully supervised approaches required significantly more training data - both in the form of labels as well as training epochs. Additionally, the TCN approach [27] uses the accelerometer data as well to achieve the state-of-the-art performance of 74.4%

| Supervision | Approach | MoF |
|---|---|---|
| Full | VGG**[21] | 7.6% |
| | IDT**[21] | 54.3% |
| | S-CNN + LSTM[21] | 66.6% |
| | TDRN[22] | 68.1% |
| | ST-CNN + Seg[21] | 72.0% |
| | TCN[27] | **73.4%** |
| None | LSTM + KNN[4] | 54.0% |
| | Ours (LSTM + AL) | **60.6%** |

Table 2: Segmentation Results on the 50 Salads dataset, at granularity '*Eval*'. **Models were intentionally reported without temporal constraints for ablative studies.

**INRIA Instructional Videos Dataset:** Finally, we evaluate our approach on the INRIA Instructional Videos dataset, which posed a significant challenge in the form of high amounts of background (noise) data. We report the F1 score for fair comparison to the other state-of-the-art approaches. As can be seen from Table 3, the proposed model outperforms the other unsupervised approach [30] by 23.3%, the weakly supervised approach [6] by 24.8% and has competitive performance to the fully supervised approaches[24, 3, 30].