Supplementary Material: Dual Perspective Network for Audio-Visual Event Localization

Varshanth Rao[1], Md Ibrahim Khalil[1], Haoda Li[1,2], Peng Dai[1], and Juwei Lu[1]

[1] Huawei Noah's Ark Lab
{varshanth.rao1, md.ibrahim.khalil, haoda.li, peng.dai, juwei.lu}@huawei.com
[2] University of Toronto, Canada

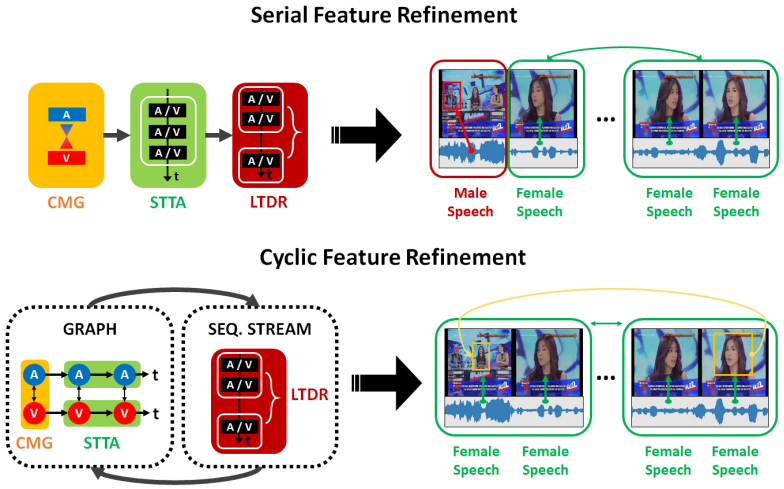
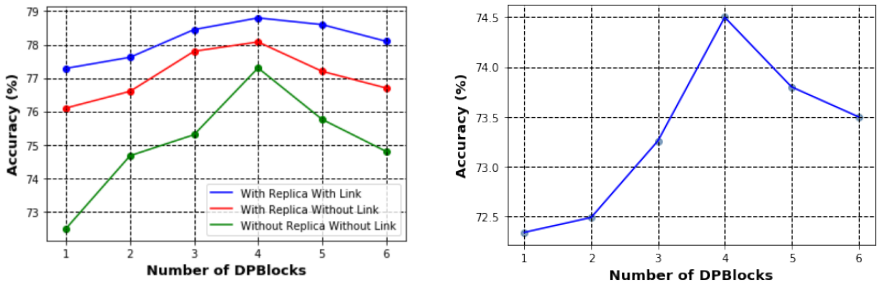


Fig. 1: An extension of Fig. 1 from the main submission. We demonstrate the advantage of cyclic feature refinement over serial feature refinement through a hypothetical example. On top, an AVEL model executes serial feature refinement. Its cross-modal guidance (CMG) module can wrongly associate the audible dialogue with the male. It can then separate that segment from that of the neighboring segments of female speech during short-term temporal association (STTA) and long term temporal dependency resolution (LTDR), resulting in the segment's erroneous prediction of male speech. During LTDR, the model learns the representation of the female speaker and refines segment features accordingly. On the bottom, during cyclic feature refinement, the LTDR output from the previous run can convey to the CMG module of the next run, that the speaker is a female. This can induce a corrective cascading effect to further modules, resulting in the segment's correct prediction of female speech.



(a) SEL task.

(b) WSEL task.

Fig. 2: Ablation study on the effect of the number of DPBlocks on the SEL and WSEL performance on the O-AVE dataset.

1 Serial vs Cyclic Refinement: Ablation on Number of DPBlocks

In this section, we investigate the influence of the number of Dual Perspective Blocks (DPBlocks) on the performance of the DPNet for the SEL and WSEL tasks on the O-AVE dataset. We vary the number of DPBlocks (B) in the range [1, 6] and plot the performance in Fig. 2. With $B = 1$ representing serial feature refinement, an increase in the number of DPBlocks increases the number of feature refinement cycles that the modal segment features go through. Hence, this study provides direct insights on the efficacy of the cyclic feature refinement process. From the plots, we observe that across all the settings, $B = 4$ yields the best performance. In general, the performance increases with the increase in B till $B = 4$, post which we observe the performance decrease. Quantitatively, this implies that cyclic feature refinement results in the production of more discriminative segment features as compared to serial feature refinement ($B = 1$) for the event localization task. Between any two RGCTs in the DPNet, there is a TCN layer which creates a barrier for a strong and direct oversmoothing effect. We hypothesize that as we increase the DPBlocks, and hence the number of RGCT layers, the individual weak oversmoothing effect accumulates and becomes strong enough to incorrectly smooth the localization predictions. Upon inspecting the localizations for $B = 6$, we corroborated the above hypothesis by confirming that event predictions for contiguous patches of segments were identical with nearly identical segment-wise event distributions from the softmax activation.

2 Feature Activation Map (FAM) Visualizations

In Fig. 3, we provide additional FAM visualizations for videos of various events to demonstrate the spatial focus performed by the model during the transformation of the segment nodes into its relational polymorphs.

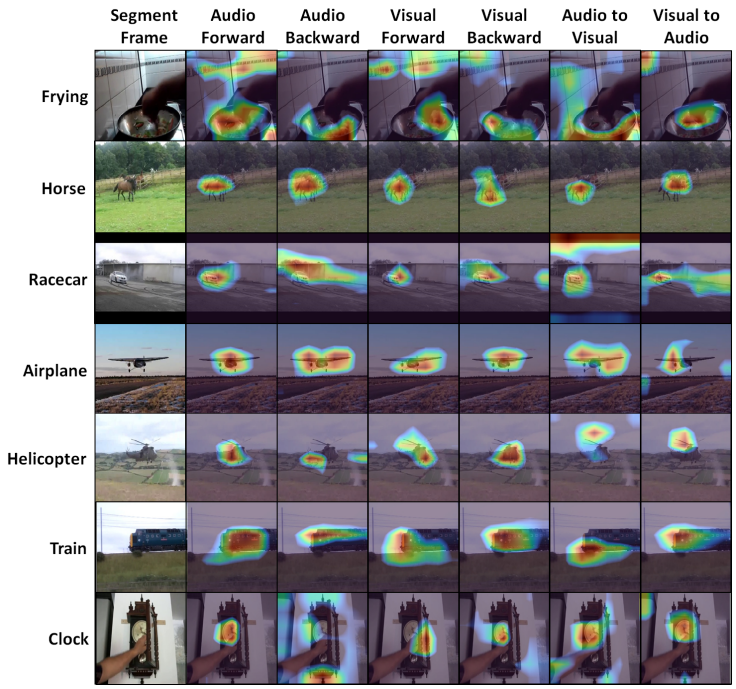
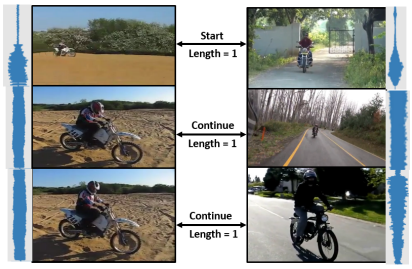


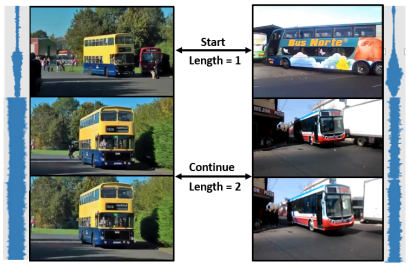
Fig. 3: Visualizations of the Feature Activation Maps from query projections of the audio and visual nodes into relational polymorphs. Each relational polymorph hones its focus onto spatial regions relevant to its semantic functionality, contributing to a rich node update.

3 Replica Visualizations

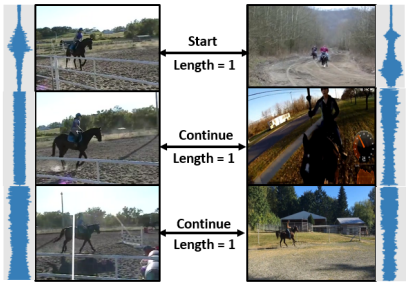
In Fig. 4, we provide audio-visual sequence visualizations of a few generated replica videos from the Replicate operation of the Replicate and Link video augmentation technique.



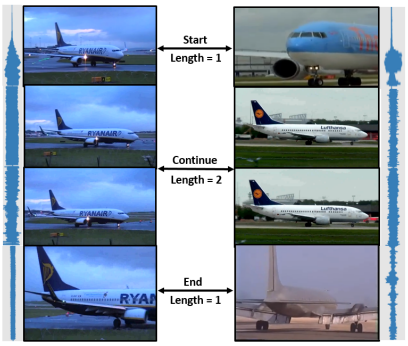
(a) Event: Motorcycle



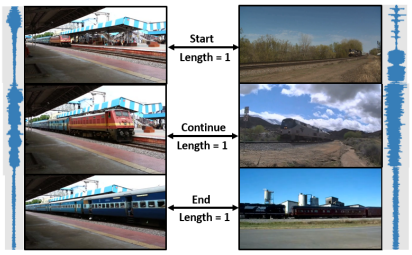
(b) Event: Bus



(c) Event: Horse



(d) Event: Aircraft



(e) Event: Train Horn

Fig. 4: Visualization of the replica videos generated by the Replicate video augmentation technique. The original video clips are plotted on the left, the corresponding sub-event sequence is conveyed in the middle, and the generated replica video clips corresponding to the sub-event sequence are plotted on the right.