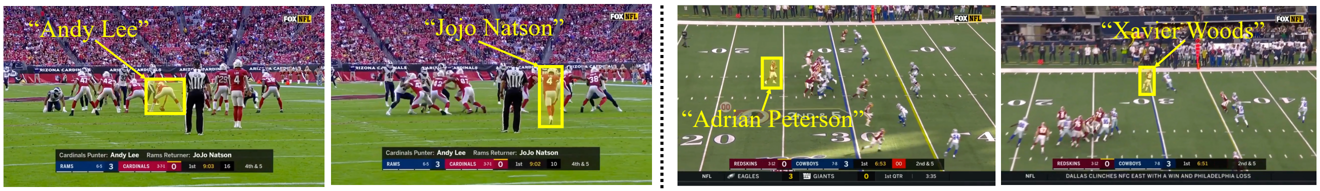
Supplementary Materials

Case study

To further investigate how the three modalities can help solve the football possession task, we randomly select two examples and show the comparisons between ground truth and the predicted results of BAPOTer.



Transcript: "…so here comes Andy Lee, the veteran Potter who has pulled off of fake is successful when I get to the Buccaneers, Jojo Natson back to receive for the Rams data gets out of 25."

Ground truth: Andy Lee, Jojo Natson
BAPOTer: Andy Lee, Jojo Natson
Ground truth: Arizona Cardinals, Los Angeles Rams
BAPOTer: Arizona Cardinals, Los Angeles Rams

(a) Correct Example

Transcript: "… Cowboys get another check away, second fumble of the Year lost by Adrian Peterson as Malcolm Smith in his second week with Dallas forced it and Xavier Woods comes away."

Ground truth: Case Keenum, Adrian Peterson, Xavier Woods
BAPOTer: Adrian Peterson, Xavier Woods
Ground truth: Washington Redskins, Dallas Cowboys
BAPOTer: Washington Redskins, Dallas Cowboys

(b) Partially Correct Example

Figure 1. Randomly selected examples from BAPO. The transcripts and corresponding video frames, ground truth and prediction results are provided. (a) is a correct prediction while (b) is a partially correct prediction. We underline the mentioned possessors in the transcript and mark them with bounding box in the video frames.

In the first example (Figure 1a), given the video, audio and text, BAPOTer correctly predicts the two players and two teams who possess the ball. However, in the second example (Figure 1b), only two possessors are correctly predicted while there are three possessors in the ground truth. The reason might be that only Adrian Peterson and Xavier Woods are mentioned in the audio and transcript. This is consistent with our experimental results that demonstrated that language is the most important modality. This example tells us that there are possessors appearing in the video but not mentioned in the audio or text. Thus, the video is sometimes a more complete depiction than the audio or transcript.