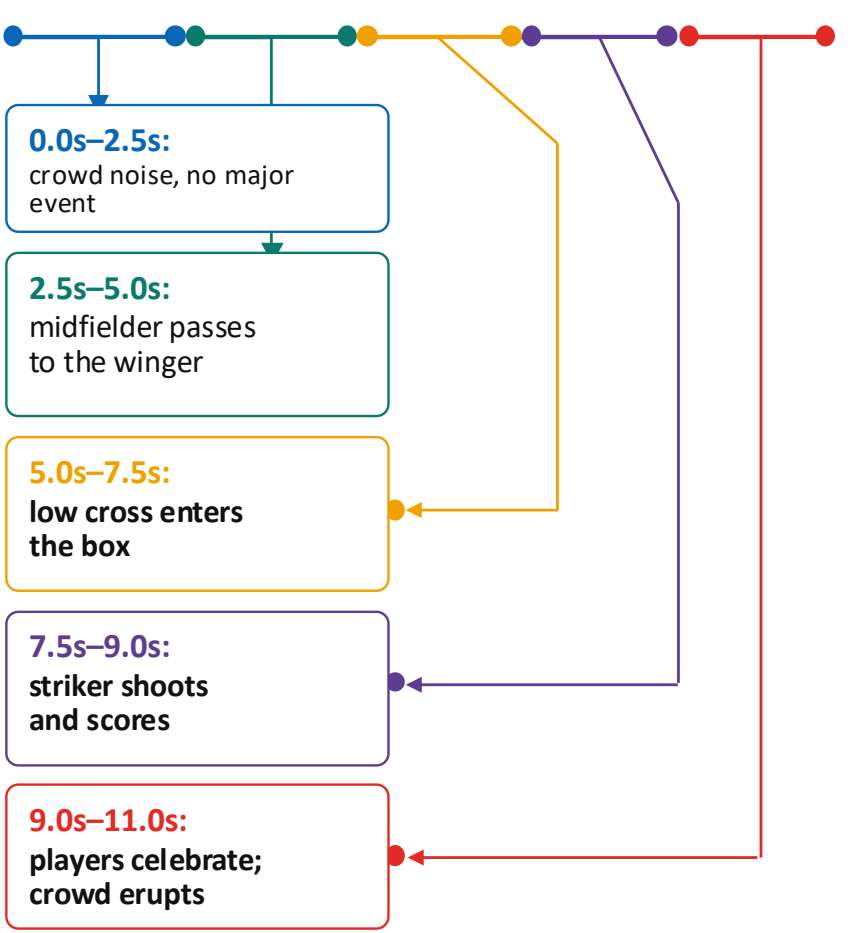
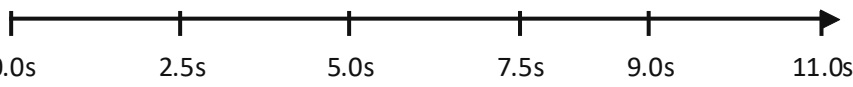


1 Detailed Video Caption Dataset

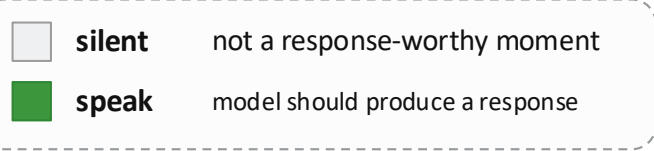
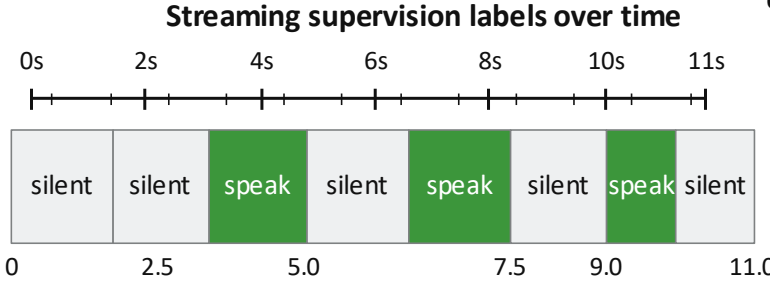
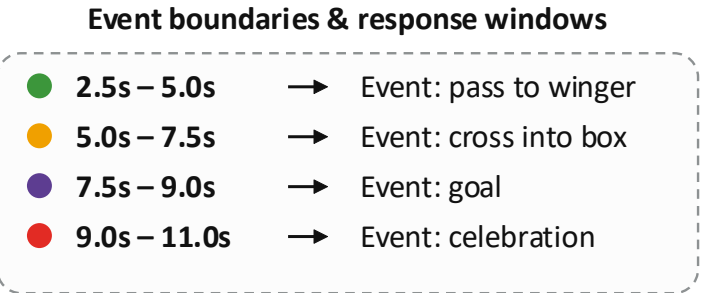
Soccer match video timeline (0 – 11.0s)



Fine-grained captions with start/end timestamps

2 Convert to Streaming Supervision

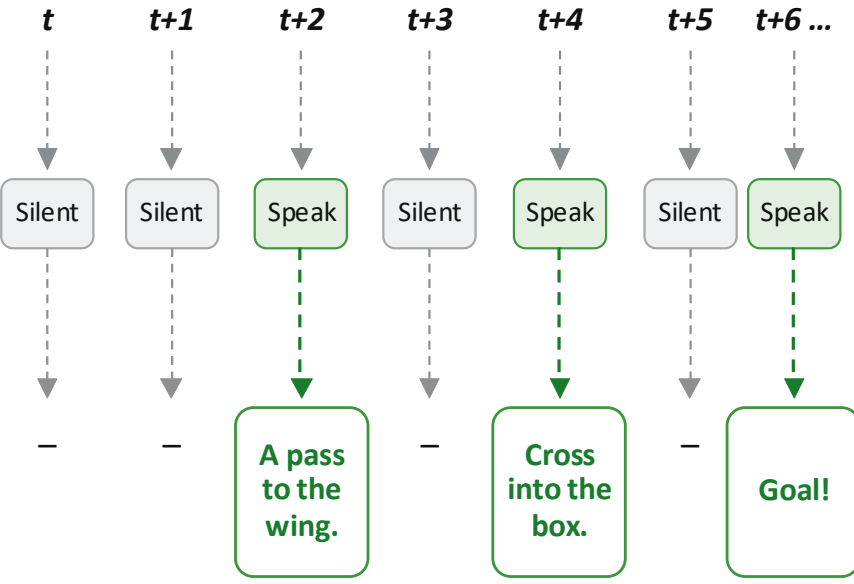
derive response-worthy events and speaking intervals



Converter turns dense caption supervision into a sequence of training targets over time.

3 Prompt + Streaming Input

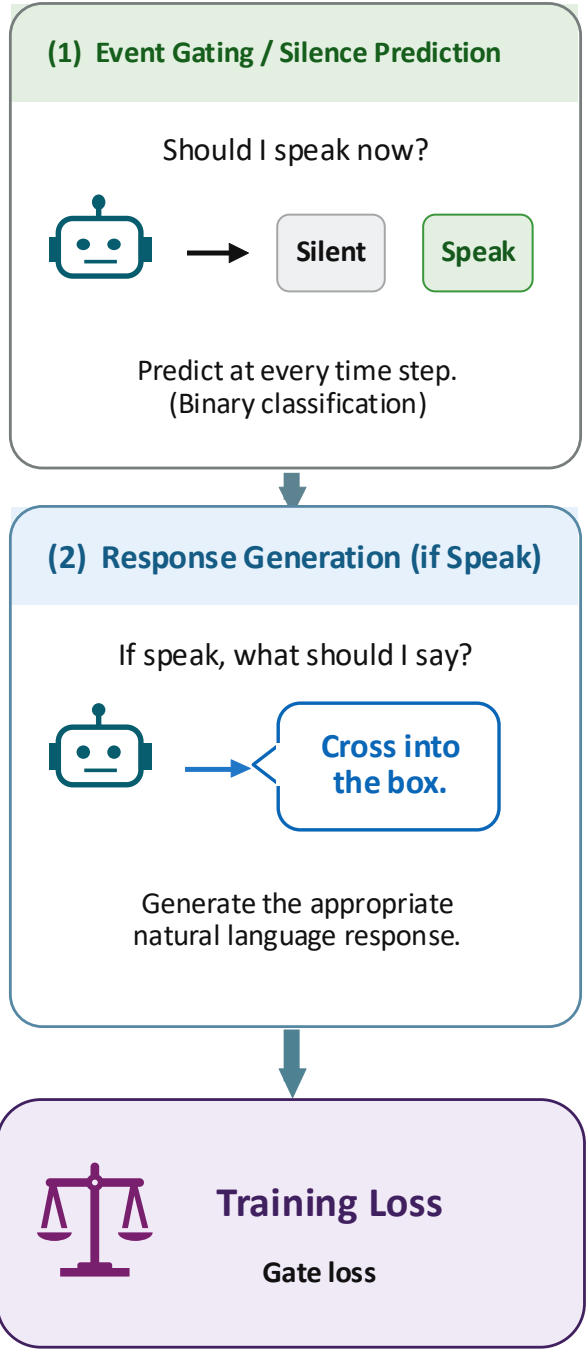
What is happening in the video?
Describe what is happening in the match.



- Model should remain **SILENT** during background segments.
- Model should **SPEAK** at events and produce the right utterance.

4 Training Objective

Model learns two behaviors jointly.



From dense temporal captions to streaming data: the model learns when to remain silent and when to respond.