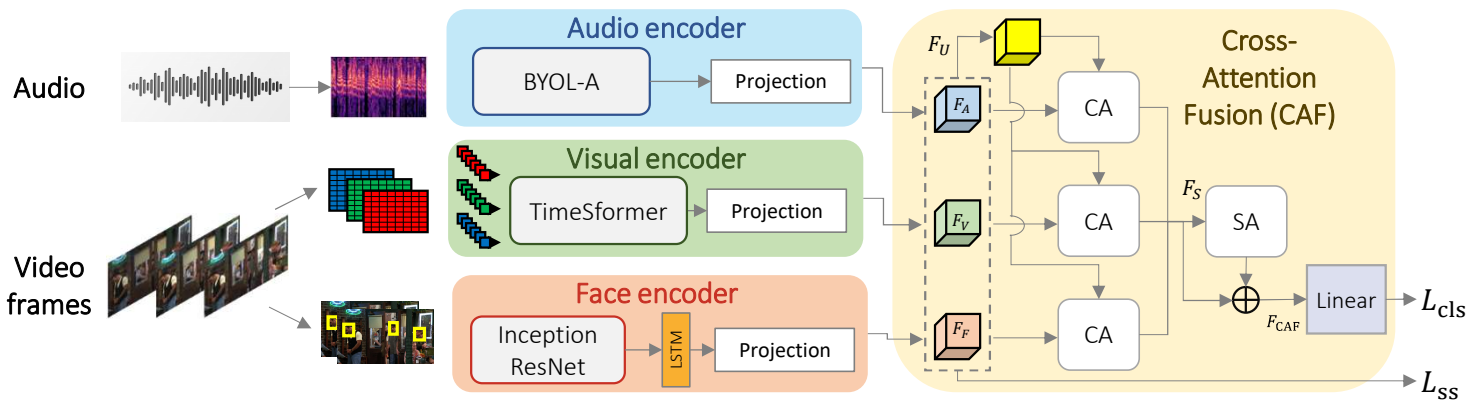
Audio
Video frames
Audio encoder
BYOL-A
Projection
Visual encoder
TimeSformer
Projection
Face encoder
Inception ResNet
LSTM
Projection
F_U
F_A
F_V
F_F
CA
CA
CA
Cross-Attention Fusion (CAF)
F_S
SA
F_{CAF}
Linear
L_{cls}
L_{ss}