Stage i

Multi-Scale STConv

Calibration

STI_GATE

Stage i+1

t

Feature Extractor

Stage i F_{stagei}

MSTF

Stage i+1 $F_{stagei+1}$

1D CNN

BiLSTM

Primary Classifier

HFEAM

Predicted Sentence

$\mathcal{L}_{ctc}$

$\mathcal{L}_{aux}$

F_{stagei} $F_{stagei+1}$

$Conv_{BR}^{s_i}$ $Conv_{BR}^{s_{i+1}}$

HA_i × × HA_{i+1}

f_i f_{i+1}

FC

AvgPool

TCN ×2

Auxiliary Classifier

Conv

ReLU

BN

MaxP

$\mathcal{L}_{aux,i+1}$ $\mathcal{L}_{aux}$

× multiplication   S sum   + addition HA_i hybrid attention module

Calibration：generates adaptive branch weights based on original features.
STI_GATE：Spatio-Temporal Interaction Gate learns an adaptive gating map on the concatenation of fused and original features to suppress background noise and enhance discriminative regions.
Multi-Scale STConv：Multi-Scale Spatio-Temporal Convolution extracts spatiotemporal features at multiple scales using parallel temporal and spatial convolutions.