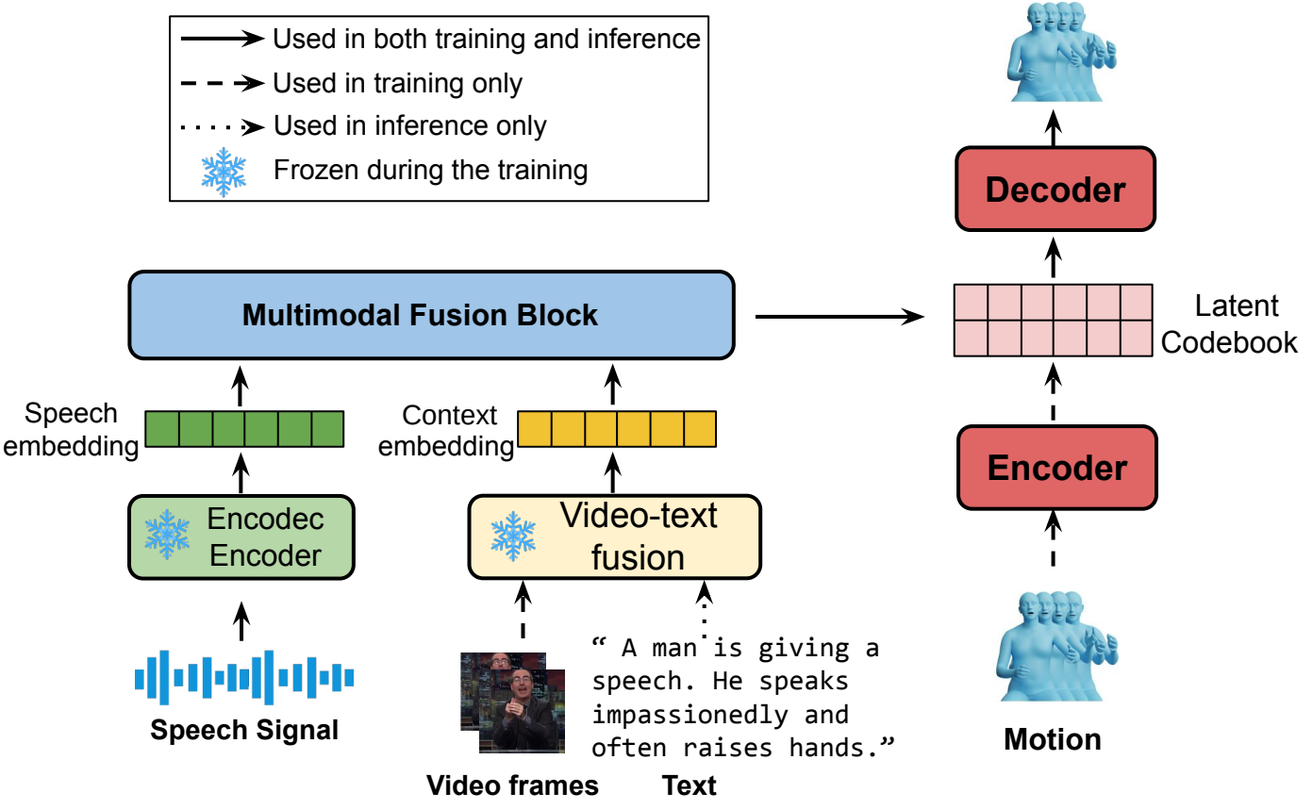
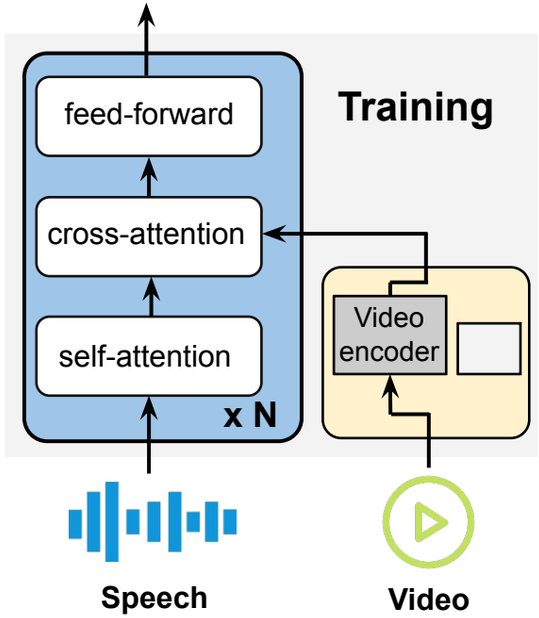
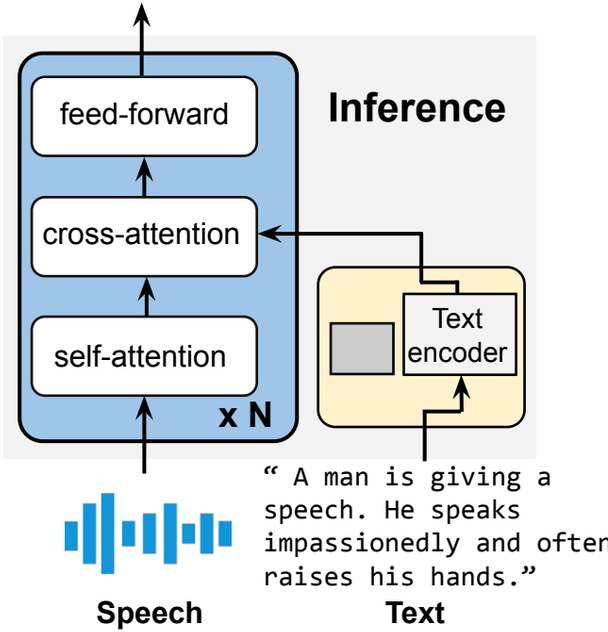

Used in both training and inference
Used in training only
Used in inference only
Frozen during the training
Multimodal Fusion Block
Speech embedding
Context embedding
Encodec Encoder
Video-text fusion
Speech Signal
Video frames
Text
“ A man is giving a speech. He speaks impassionedly and often raises hands.”
Decoder
Latent Codebook
Encoder
Motion

(a)


feed-forward
cross-attention
self-attention
x N
Training
Video encoder
Speech
Video


feed-forward
cross-attention
self-attention
x N
Inference
Text encoder
Speech
“ A man is giving a speech. He speaks impassionedly and often raises his hands.”
Text

(b)