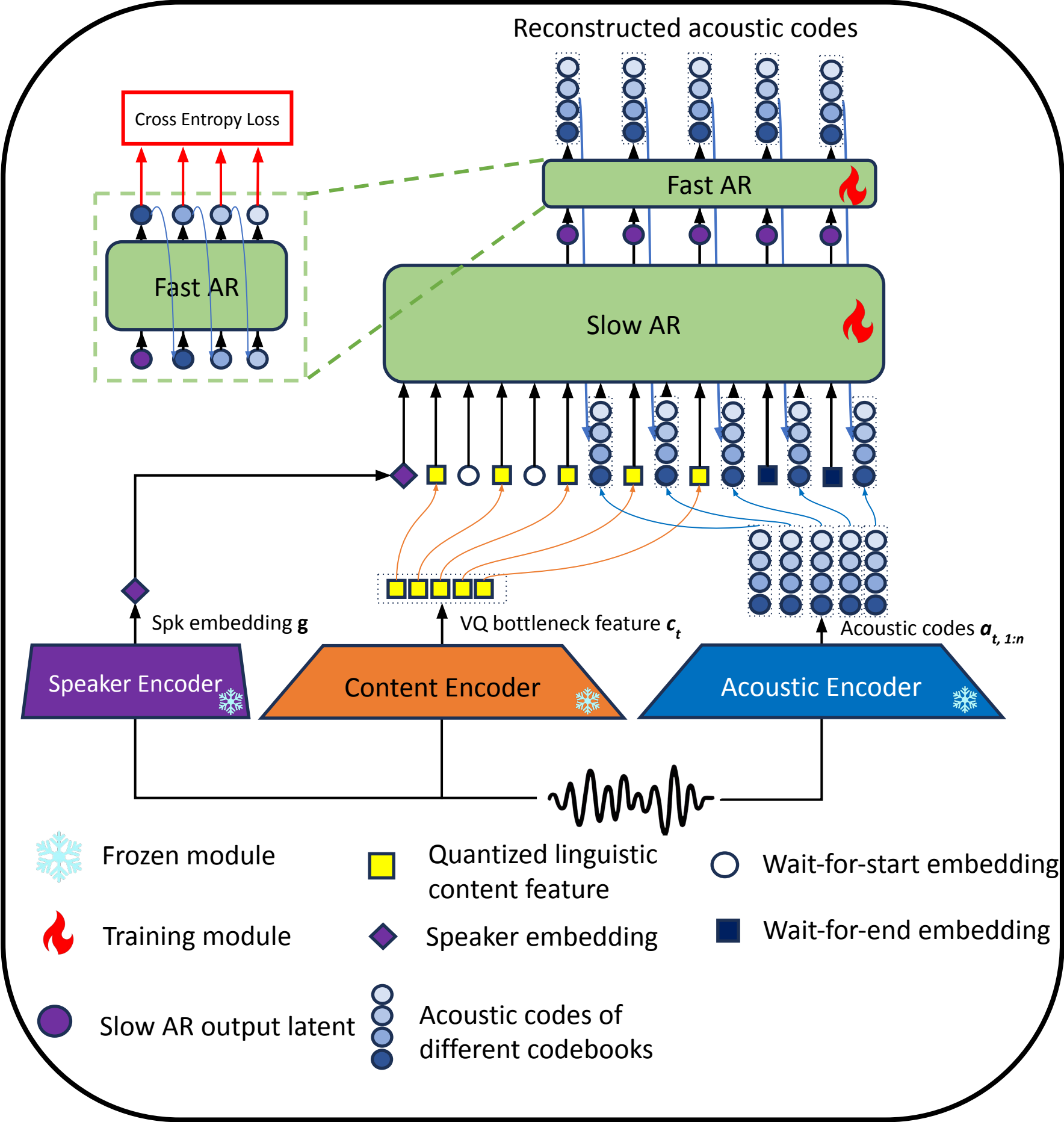
Reconstructed acoustic codes
Cross Entropy Loss
Fast AR
Fast AR
Slow AR
Spk embedding g
VQ bottleneck feature c_t
Acoustic codes $a_{t,1:n}$
Speaker Encoder
Content Encoder
Acoustic Encoder
Frozen module
Training module
Slow AR output latent
Quantized linguistic content feature
Speaker embedding
Acoustic codes of different codebooks
Wait-for-start embedding
Wait-for-end embedding