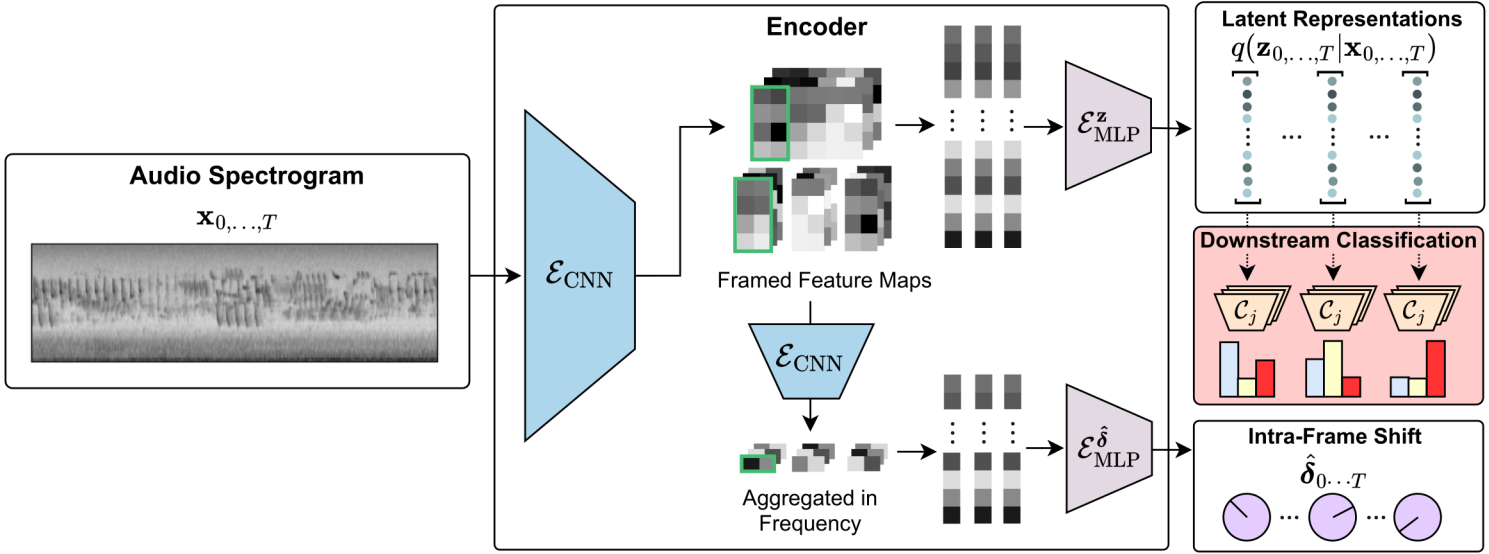
Audio Spectrogram
$\mathbf{x}_{0,\ldots,T}$
Encoder
$\mathcal{E}_{\text{CNN}}$
Framed Feature Maps
$\mathcal{E}^{\mathbf{z}}_{\text{MLP}}$
$\mathcal{E}_{\text{CNN}}$
Aggregated in Frequency
$\mathcal{E}^{\hat{\boldsymbol{\delta}}}_{\text{MLP}}$
Latent Representations
$q(\mathbf{z}_{0,\ldots,T}|\mathbf{x}_{0,\ldots,T})$
Downstream Classification
$\mathcal{C}_j$
$\mathcal{C}_j$
$\mathcal{C}_j$
Intra-Frame Shift
$\hat{\boldsymbol{\delta}}_{0\ldots T}$