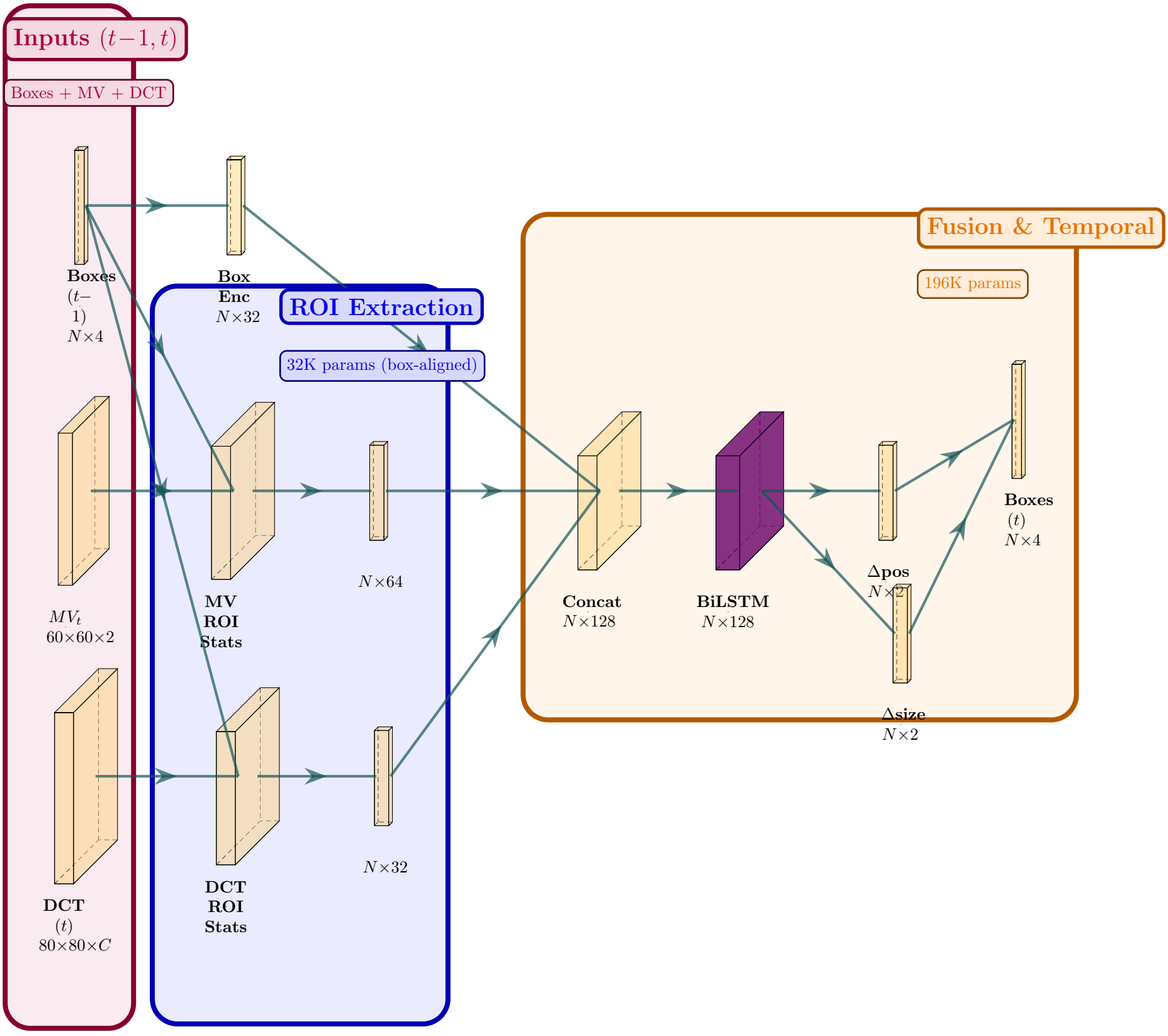
Inputs $(t-1, t)$
Boxes + MV + DCT
Boxes
$(t-1)$
$N\times4$
Box Enc
$N\times32$
ROI Extraction
32K params (box-aligned)
MV_t
$60\times60\times2$
MV ROI Stats
$N\times64$
DCT
(t)
$80\times80\times C$
DCT ROI Stats
$N\times32$
Fusion & Temporal
196K params
Concat
$N\times128$
BiLSTM
$N\times128$
Δpos
$N\times2$
Δsize
$N\times2$
Boxes
(t)
$N\times4$