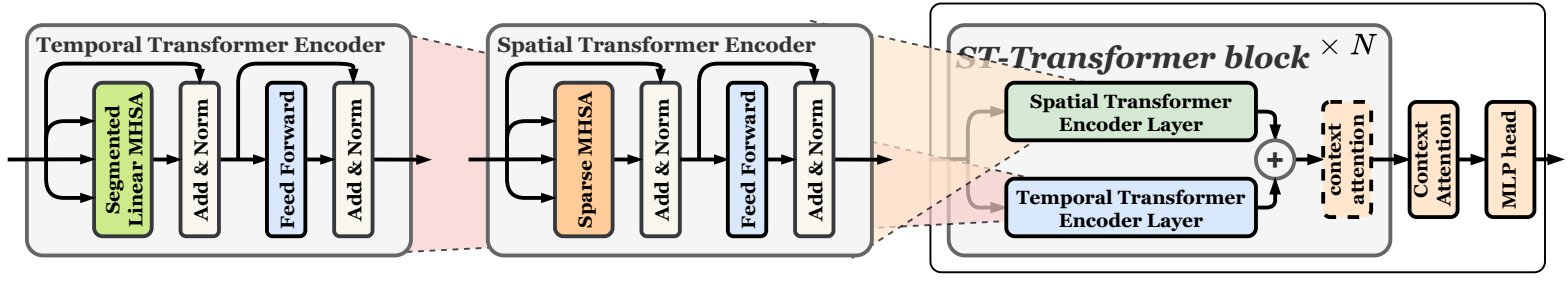

(a) Temporal Transformer Encoder Layer: the attention module is a Segmented Linear MHSA

(b) Spatial Transformer Encoder Layer: The attention module is a Sparse MHSA

(c) The architecture of our model: consists of N *ST*-Transformer block