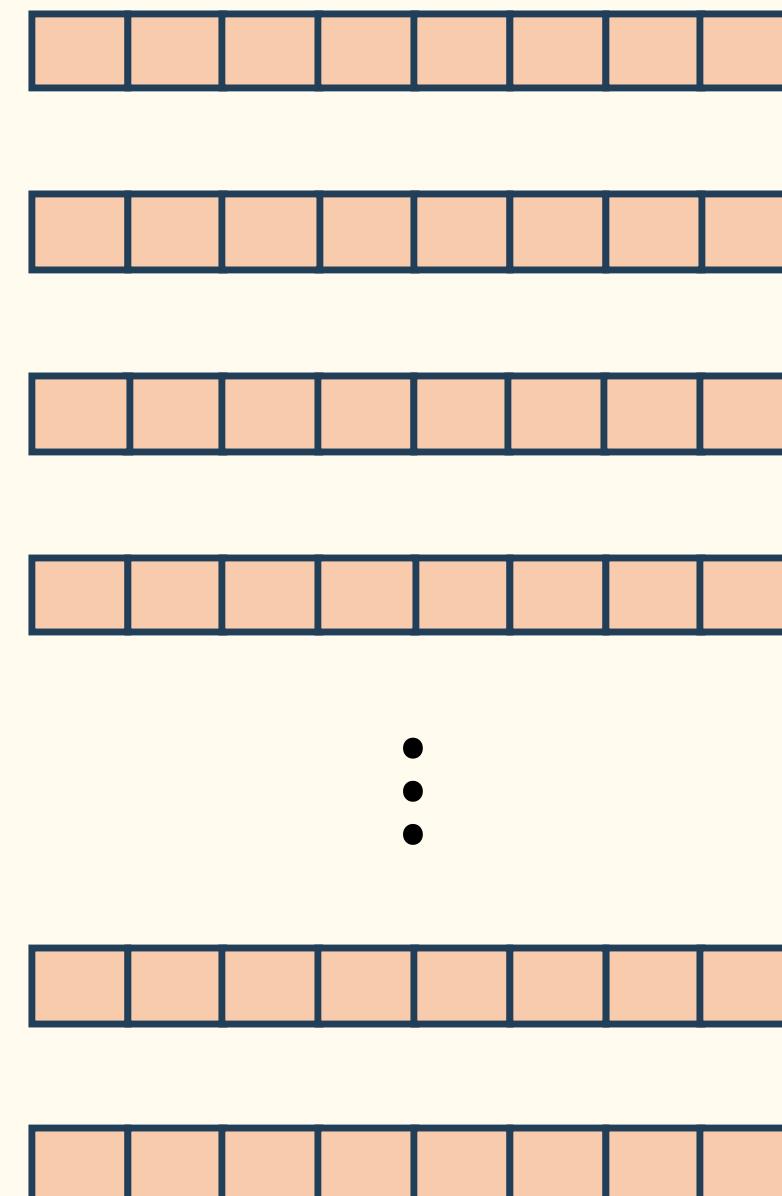
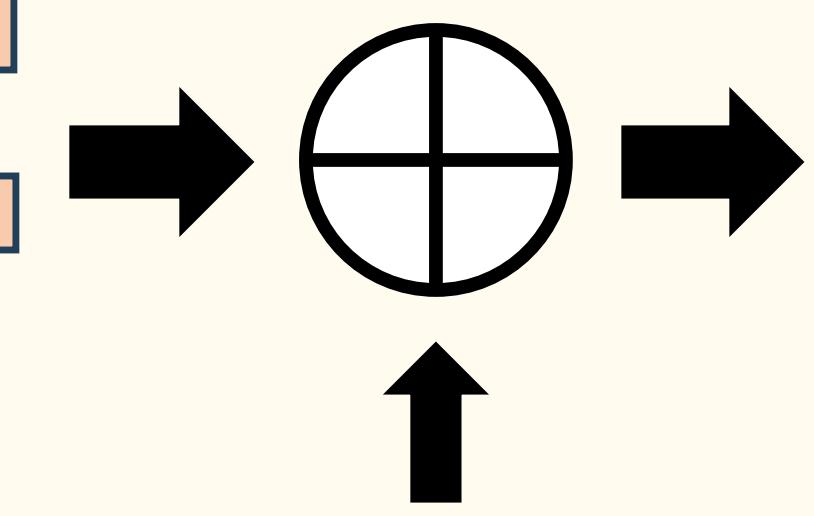


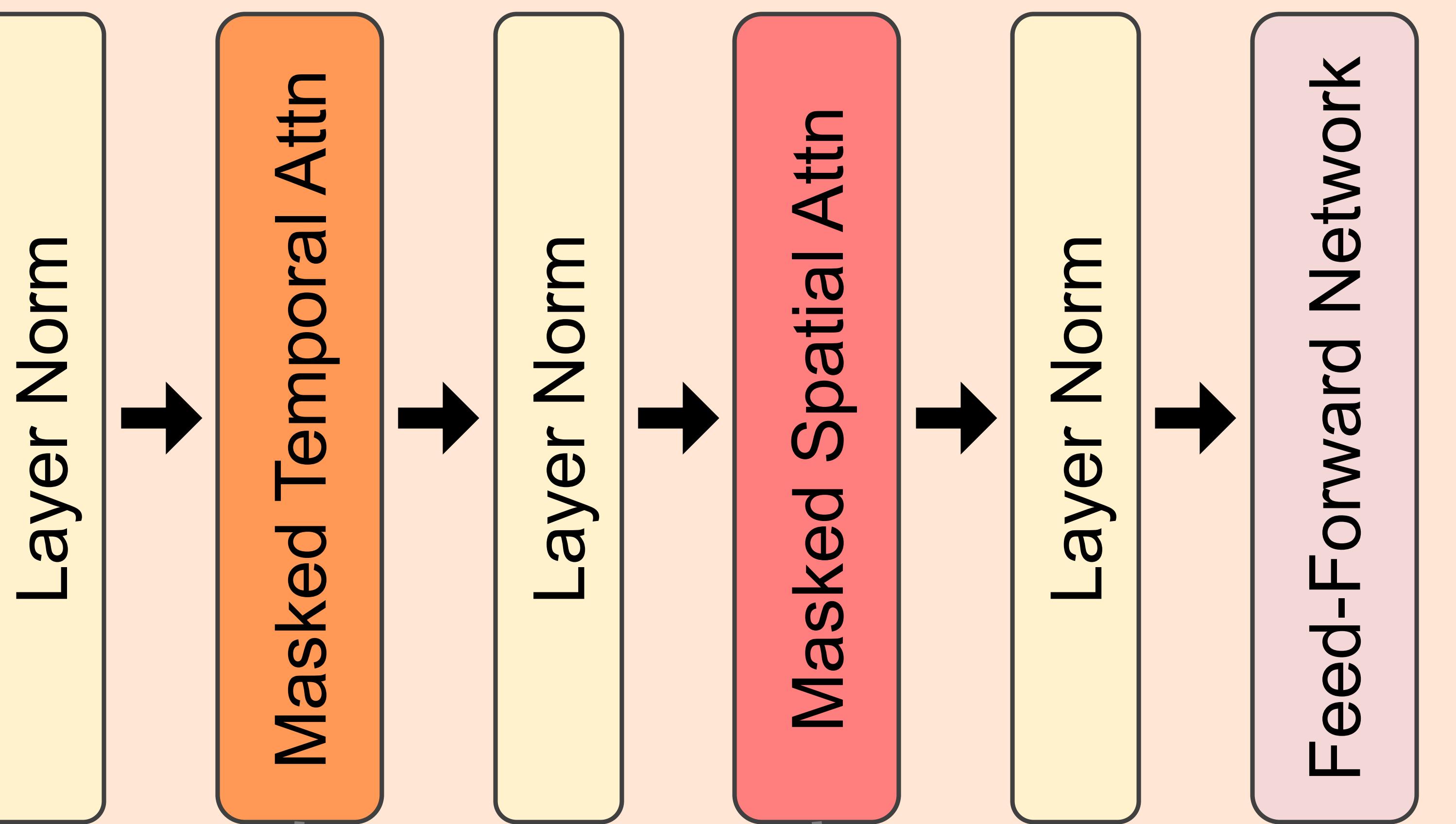
MSTA-based Feature Extractor



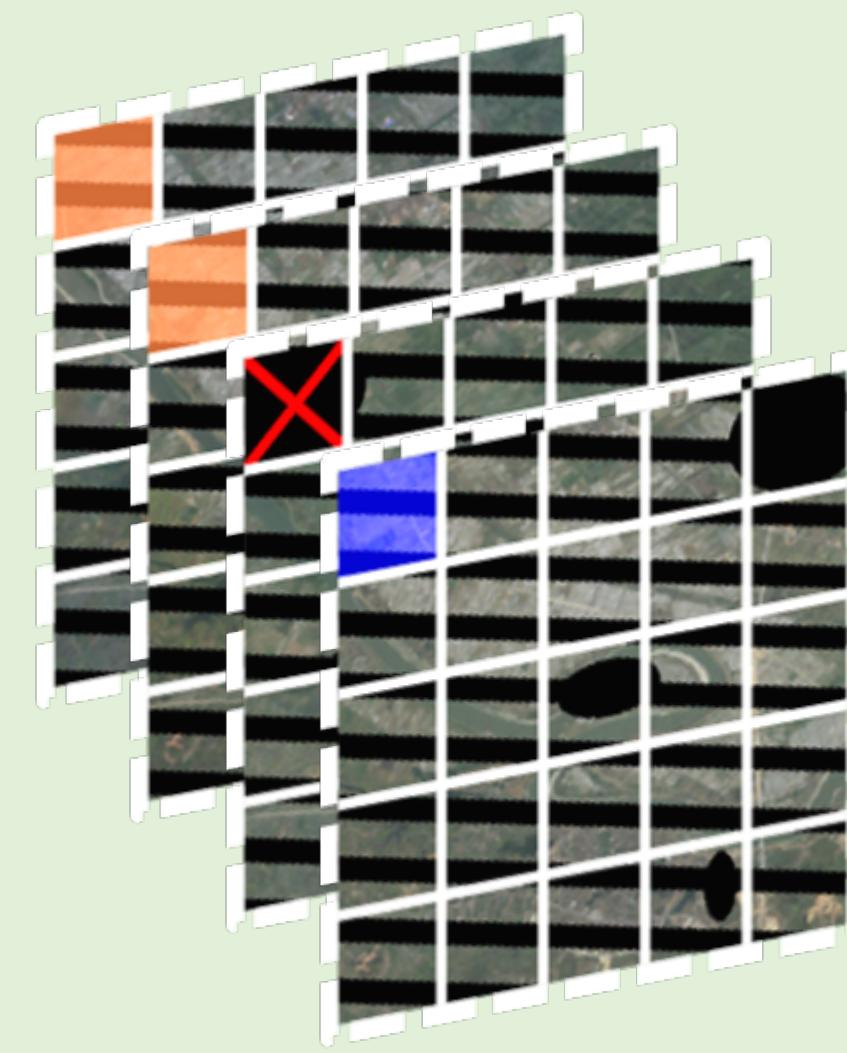
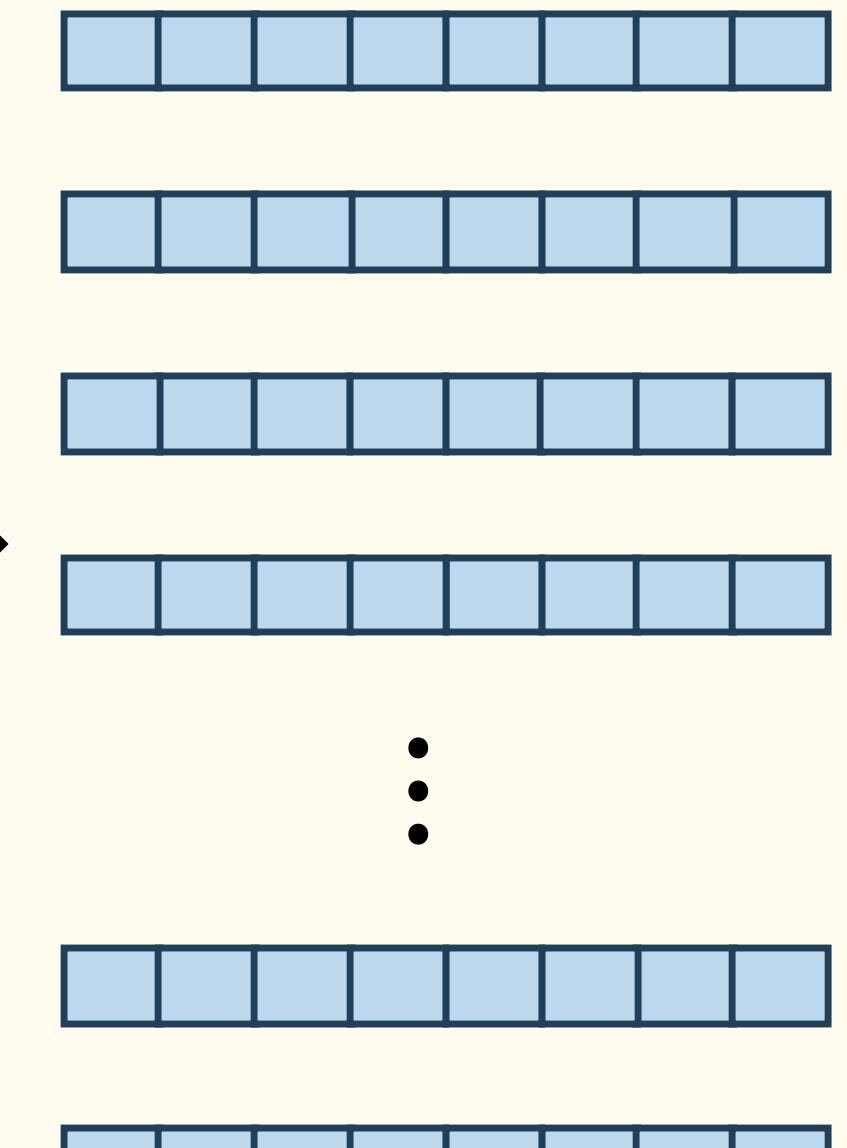
Positional
Encoding



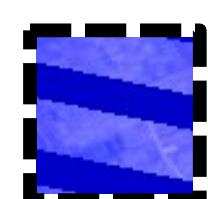
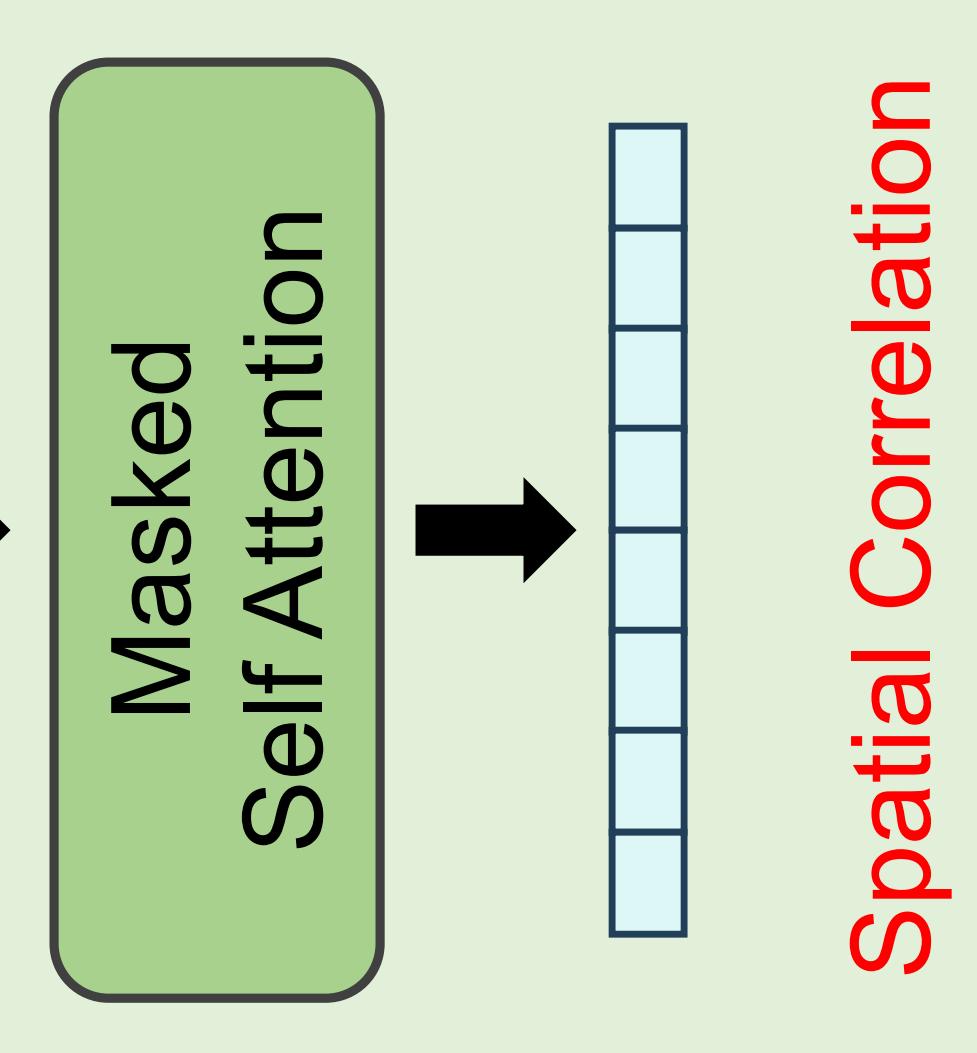
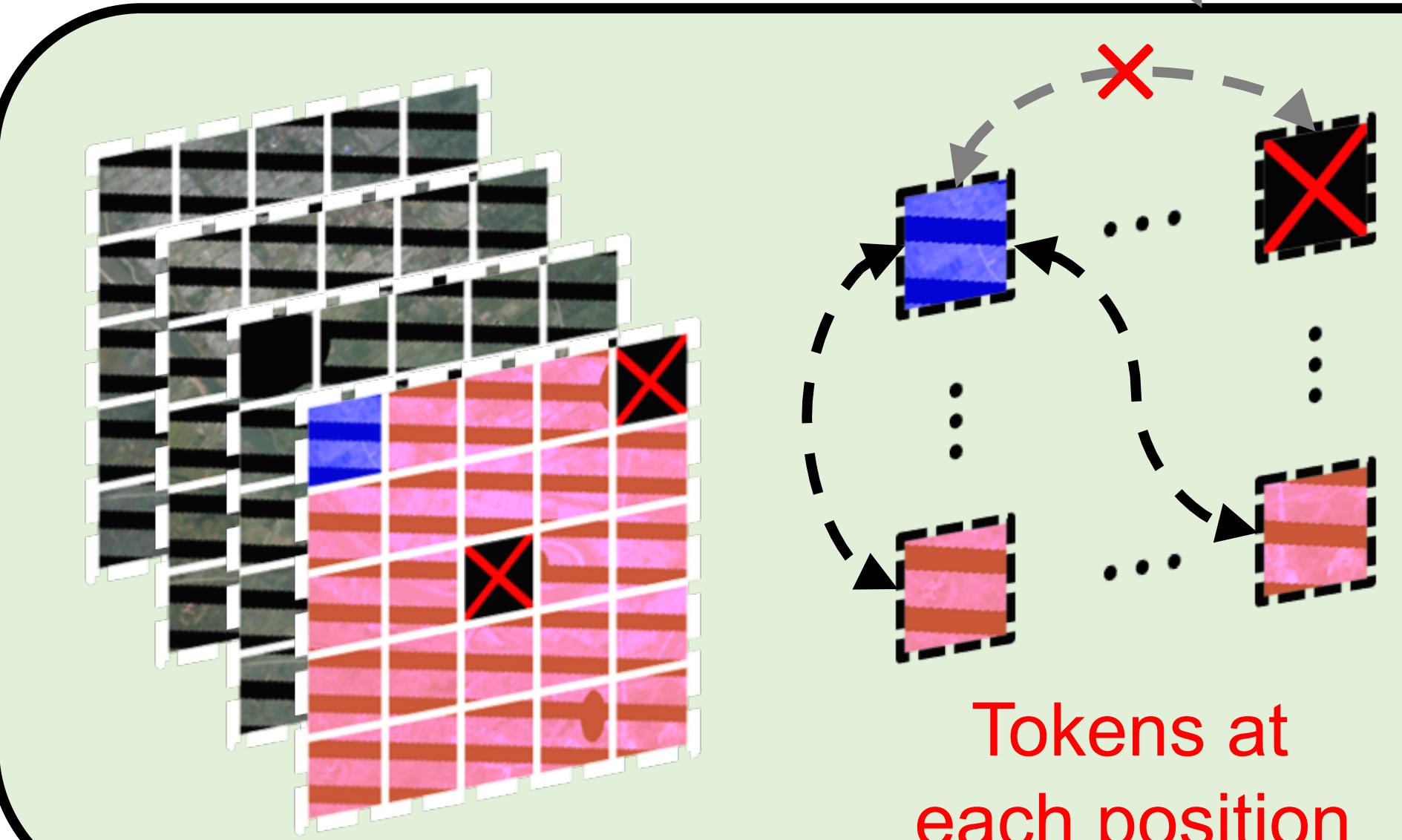
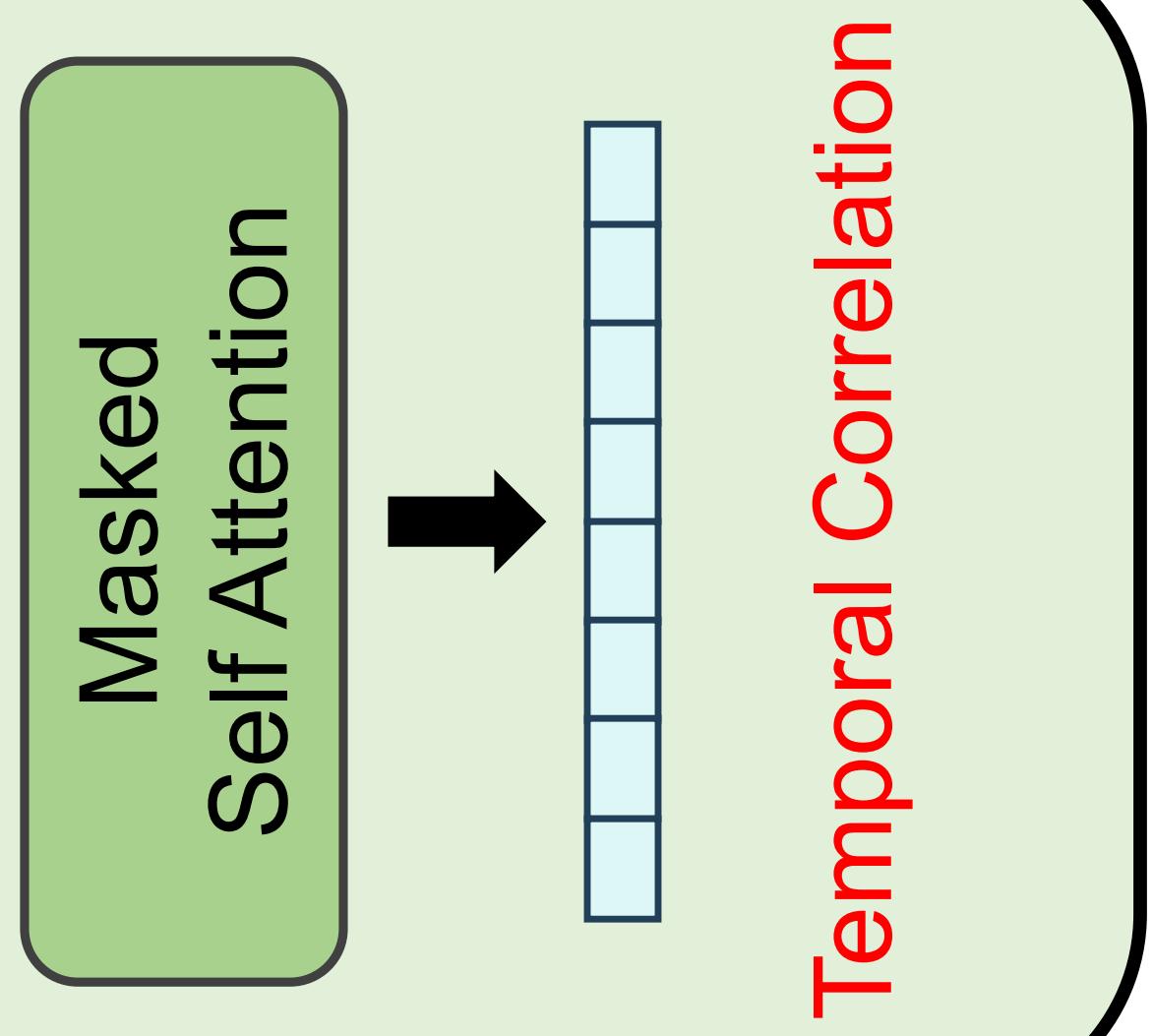
Masked Spatial-Temporal Attention



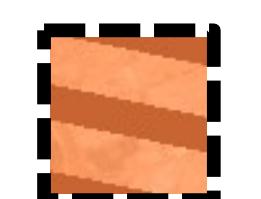
$\times L$



Tokens at each time



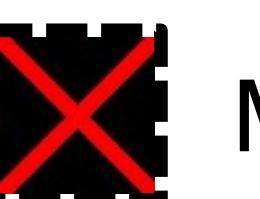
Target Patch



Temporal Auxiliary



Spatial Auxiliary



Masked Patch



Output of Target