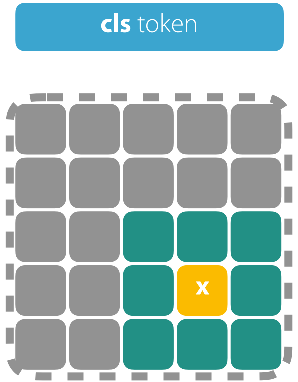
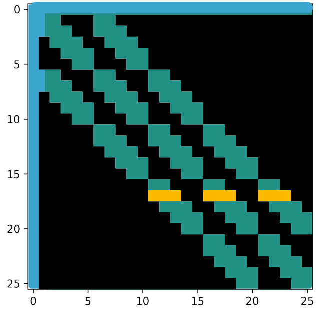
Modify attention with block mask *M*

query $\mathbf{x}_i$, the entries of *M* are set to minus ∞ all tokens outside $\mathcal{N}_i$ except the **[cls] token,** and **zero otherwise**



block mask *M*



Local attention: fast + inductive bias

$$\texttt{Attention}(Q,K,V) = \texttt{softmax}\left(\frac{QK^\top}{\sqrt{D}} + M\right)V$$