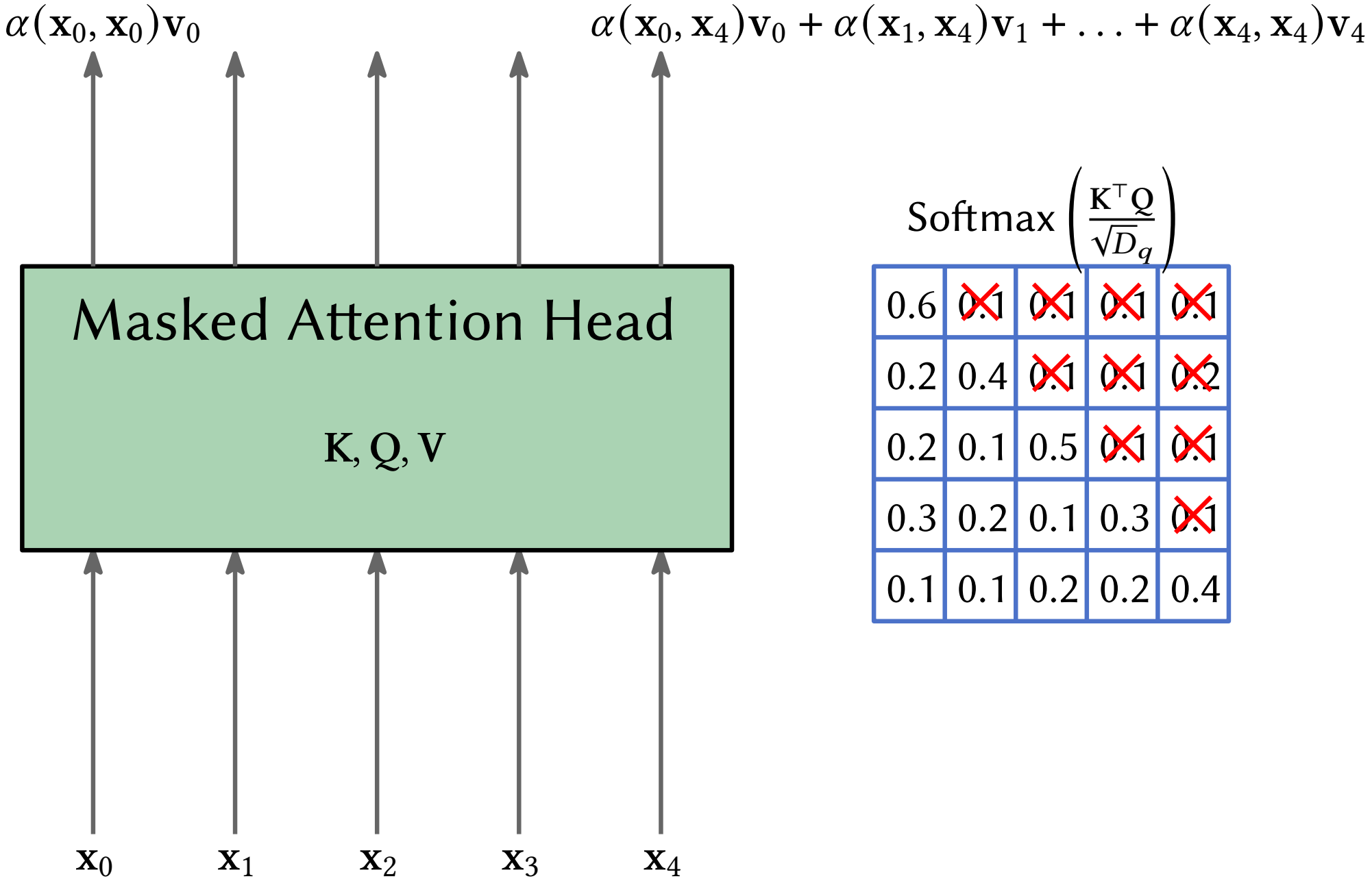
$\alpha(\mathbf{x}_0, \mathbf{x}_0)\mathbf{v}_0$
$\alpha(\mathbf{x}_0, \mathbf{x}_4)\mathbf{v}_0 + \alpha(\mathbf{x}_1, \mathbf{x}_4)\mathbf{v}_1 + \ldots + \alpha(\mathbf{x}_4, \mathbf{x}_4)\mathbf{v}_4$
Masked Attention Head
K, Q, V
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_4$
Softmax $\left(\frac{\mathbf{K}^\top \mathbf{Q}}{\sqrt{D_q}}\right)$
0.6 0.1 0.1 0.1 0.1
0.2 0.4 0.1 0.1 0.2
0.2 0.1 0.5 0.1 0.1
0.3 0.2 0.1 0.3 0.1
0.1 0.1 0.2 0.2 0.4