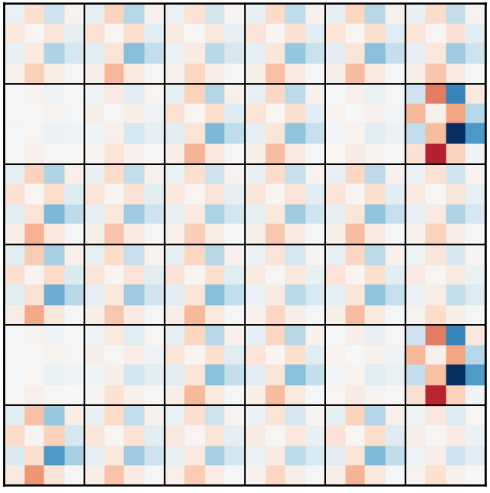
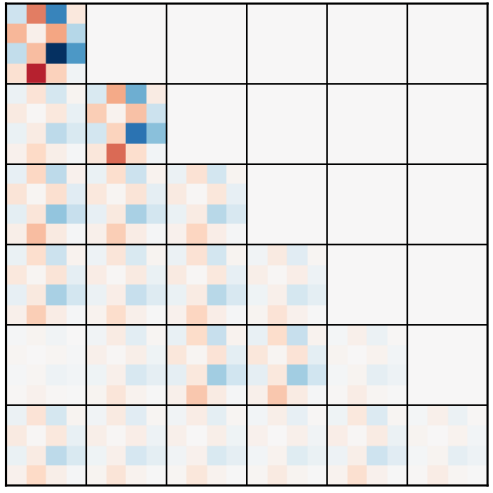
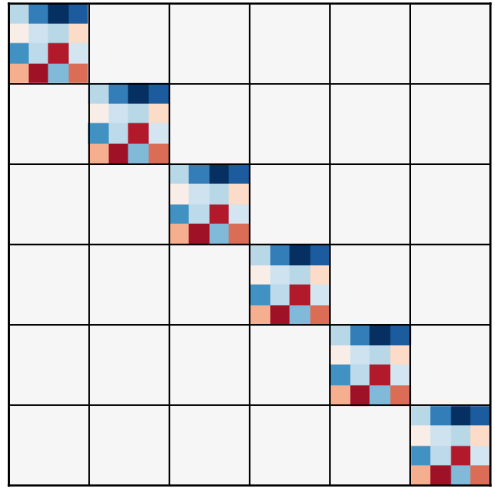

self-attention: block (n, m) = $a_{nm}\mathbf{W}^{(v)\mathrm{T}}$


masked self-attention


position-wise layer: shared diagonal blocks