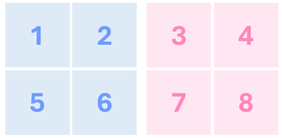
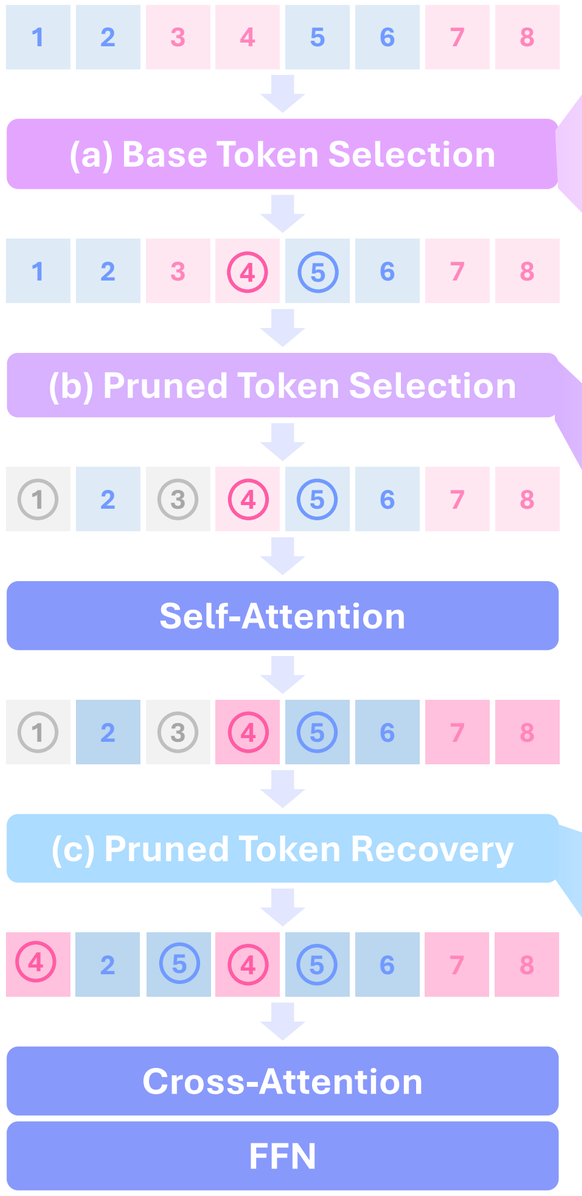
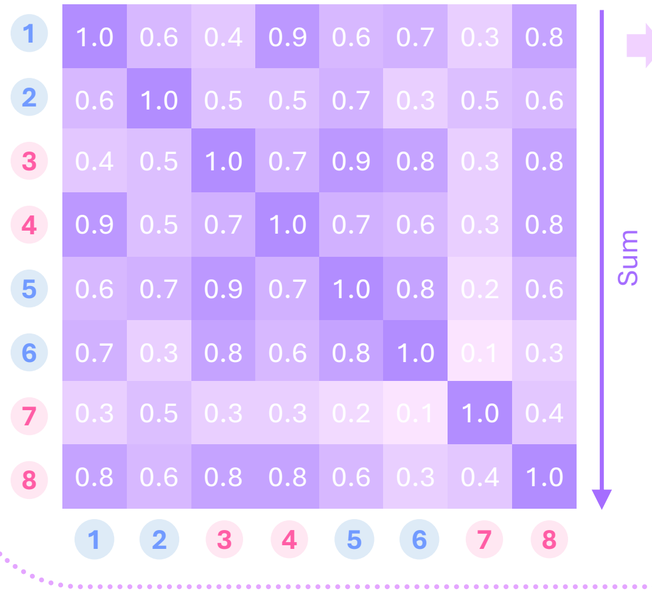
Attention Layer with SiTo

Partition the Image into patches of size $s \times s$

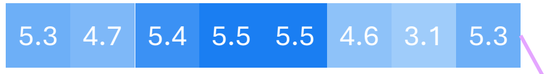




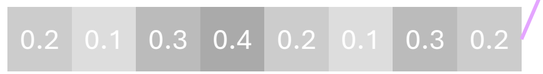
① Compute the Cosine Similarity $CosSim$ between N tokens.



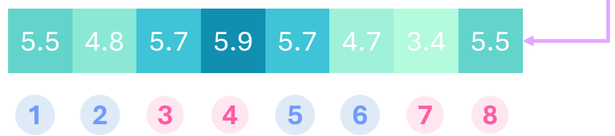
② $SimScore = Sum\ (CosSim, dim = 1)$



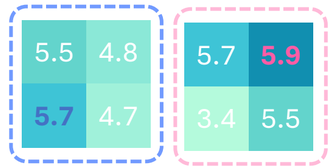
③ Add the $Guassian\ Noise$



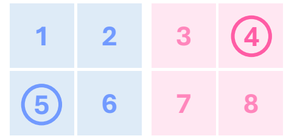
④ $Noise\ SimScore = SimScore + Noise$



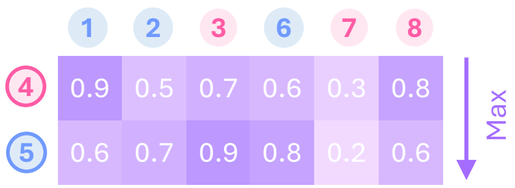
⑤ Find the maximum value in each patch.



⑥ Get the **Base Tokens** from each patch (○ ○).



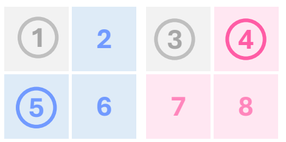
① Get the Similarity between the **Base Tokens** and others.



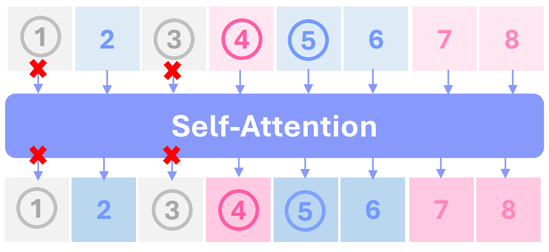
② Fine the maximum value and corresponding index.

③ Select the Top K Values.

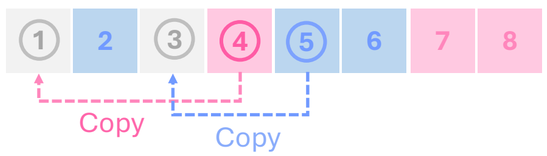
④ Get the **Pruned Tokens** (○).



① Get the output of the attention layer.



② Recover the pruned tokens from base tokens.



③ Get the final outputs.