Lab Problem: CUDA Self-Attention

Implement a CUDA program to compute a simplified self-attention operation using the Q, K, and V matrices.
Given Q, K, and V, perform the following steps:

- Transpose K to obtain K^T.
- Compute $Q \times K^T$.
- Apply row-wise Softmax.
- Compute Softmax × V to obtain the output.

The Softmax operation is:

$$Pij = \exp(Sij - Mi) / \Sigma j \exp(Sij - Mi)$$

where $S = QK^T$ and Mi is the maximum value in row i.

CUDA Requirements:
Implement a CUDA kernel for matrix transpose.
Implement a CUDA kernel for $Q \times K^T$.
Implement a CUDA kernel for row-wise Softmax using a reduction.
Implement a CUDA kernel for Softmax × V.
.

Input

Define N. Generate the matrices using sequential values for easy verification.
Small test case:
Q = [[1, 2], [3, 4]]
K = [[5, 6], [7, 8]]
V = [[1, 2], [3, 4]]

Output

K^T:
5 7
6 8

QK^T:
17 23
39 53

Softmax:
0.00247 0.99753
0.00000083 0.99999917

Output :
2.995 3.995
3.000 4.000