Self attention

$X \in \mathbb{R}^{N \times S \times W}$

where N = Batch size

S = seq length

W = width of embedding dimension

$W_{QKV} \in \mathbb{R}^{3W \times W}$

where $3W$ is QKV rows

W is width.

$x @ W_{QKV}^T \in \mathbb{R}^{N \times S \times 3W}$

L_S split info

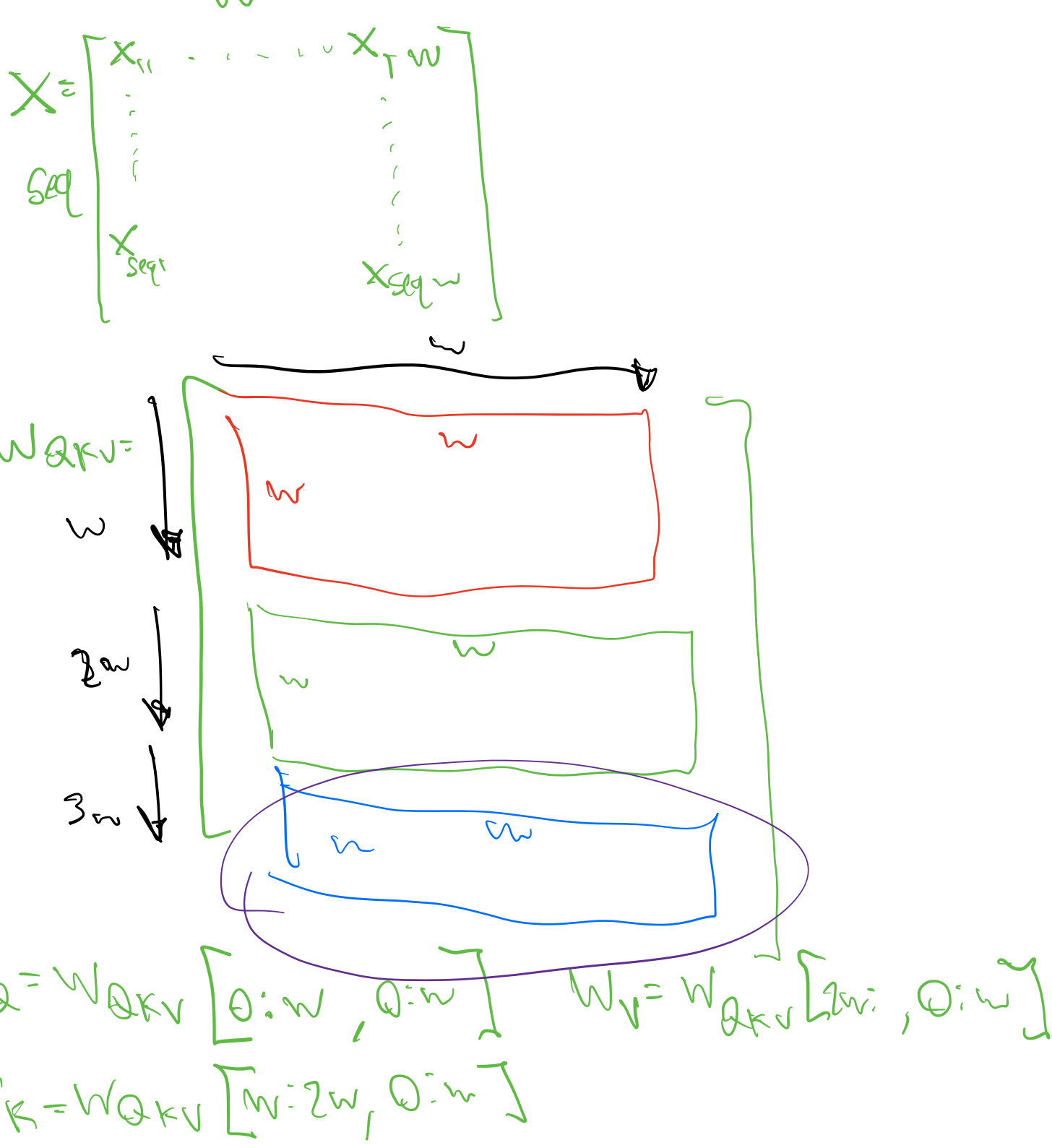
$Q \in \mathbb{R}^{N \times S \times W}$,

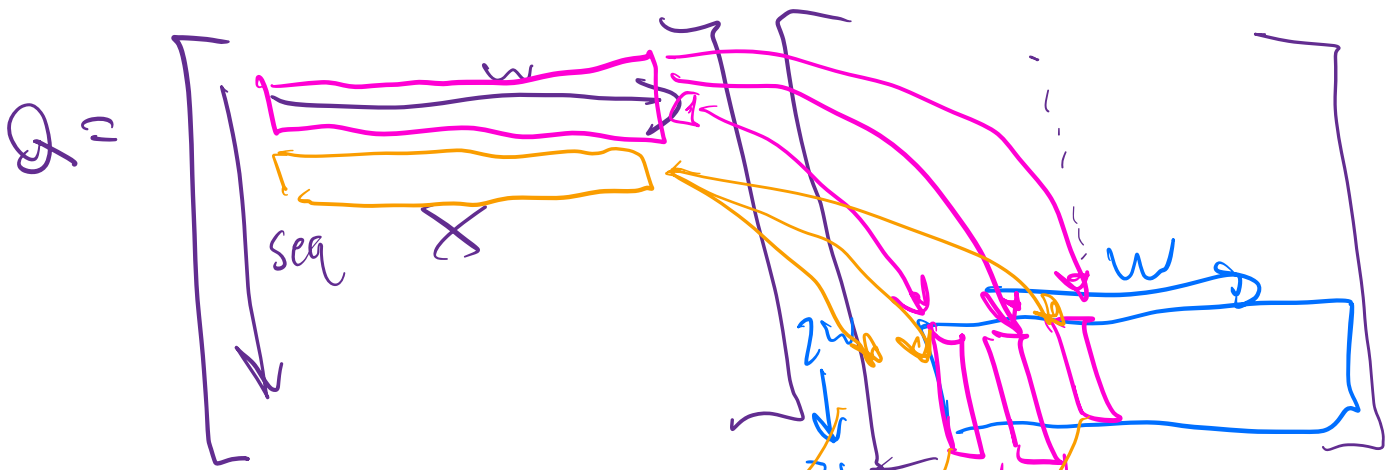
$K \in \mathbb{R}^{N \times S \times W}$,

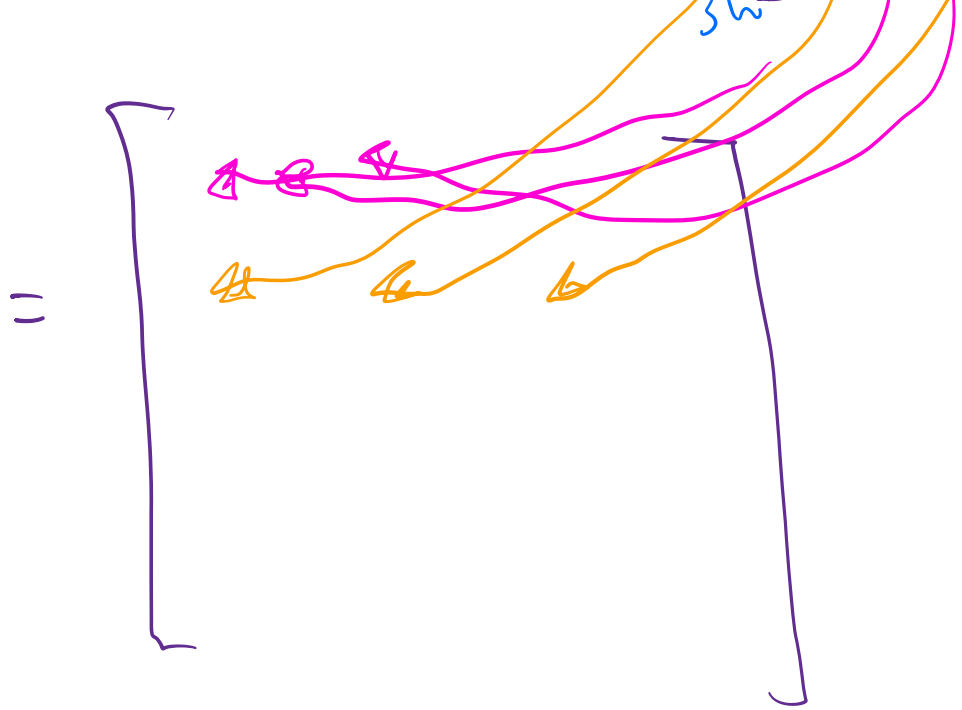
$V \in \mathbb{R}^{N \times S \times W}$

Indexing Logic:

Just focus on (S, W) from X and $(3W, W)$ from W_{QKV}


X =
seq
W
W_QKV =
W
2W
3W
W_Q = W_QKV [0:W, 0:W]
W_V = W_QKV [2W:, 0:W]
W_K = W_QKV [W:2W, 0:W]


Q =
seq
W
2W
W



$$Q_{ij} = \sum_k X_{ik} * X_{kj}$$

$k \in [0, w-1]$ or $[w, 2w-1]$ or $[2w, 3w-1]$

$i \in [0, seq]$

$j \in [0, w-1]$

```
for k in range(w):
    Q_ij += X[i, k] * w[k, j]
    K_ij += X[i, k] * w[w+k, j]
    V_ij += X[i, k * w[2w+k, j]
```

```
for i → seq:
    for j → w:
        for k → w:
            Q[i,j] += x[i,k] w[k,j]
            K[i,j] += x[i,k] w[w+k,j]
            V[i,j] += x[i,k] w[2w+k,j]
```

TVM Notation

```
for i,j,k in T.grid(seq, w, w):
    vi,vj,vk = T.axis.remap("SSS", [seq, w, w])
        Q[vi,vj] += X[vi,vk] * w[vk,vj]
        K[vi,vj] += x[vi,vk] * w[w+vk,vj]
        V[vi,vj] += x[vi,vk] * w[2w+vk,vj]
```