

(Seq-len, Seq-len) = (4, 4)
 (Block-size-Q, Block-size-KV) = (2, 2)

$$P = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0.1 & 0.2 & 0 & 0 \\ 0.1 & 0.2 & 0.3 & 0 \\ 0.1 & 0.2 & 0.3 & 0.4 \end{bmatrix}$$

$$L = \begin{bmatrix} .1 \\ .3 \\ .6 \\ 1.0 \end{bmatrix}$$

(Seq-len, Head-dim) = (4, 4)
 (Block-size-KV, head-dim) = (2, 4)

$$V = \begin{bmatrix} .1 & .2 & .3 & .4 \\ .2 & .1 & .2 & .3 \\ .3 & .2 & .1 & .2 \\ .4 & .3 & .2 & .1 \end{bmatrix}$$

Traditional Attention Ordering

$$P_{00} \div L_0 = \begin{bmatrix} 0.1 & 0 \\ 0.1 & 0.2 \end{bmatrix} \div \begin{bmatrix} .1 \\ .3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ .3 & .6 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ .3 & .6 \end{bmatrix} \begin{bmatrix} .1 & .2 & .3 & .4 \\ .2 & .1 & .2 & .3 \end{bmatrix} = \begin{bmatrix} .1 & .2 & .3 & .4 \\ .16 & .13 & .23 & .3 \end{bmatrix}$$

Flash-Attention rearranged softmax denominator

$$O_{00} = P_{00} \odot V_0 = \begin{bmatrix} 0.1 & 0 \\ 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} .1 & .2 & .3 & .4 \\ .2 & .1 & .2 & .3 \end{bmatrix} = \begin{bmatrix} .01 & .02 & .03 & .04 \\ .05 & .04 & .07 & .1 \end{bmatrix}$$

$$O_{00} / L_0 = \begin{bmatrix} .01 & .02 & .03 & .04 \\ .05 & .04 & .07 & .1 \end{bmatrix} \div \begin{bmatrix} .1 \\ .3 \end{bmatrix} = \begin{bmatrix} .1 & .2 & .3 & .4 \\ .16 & .13 & .23 & .3 \end{bmatrix}$$

Then remember that this is still not equal to the top two rows of the final O since we still need the for loop in `_attn_fwd_inner()` to accumulate over rows of V .