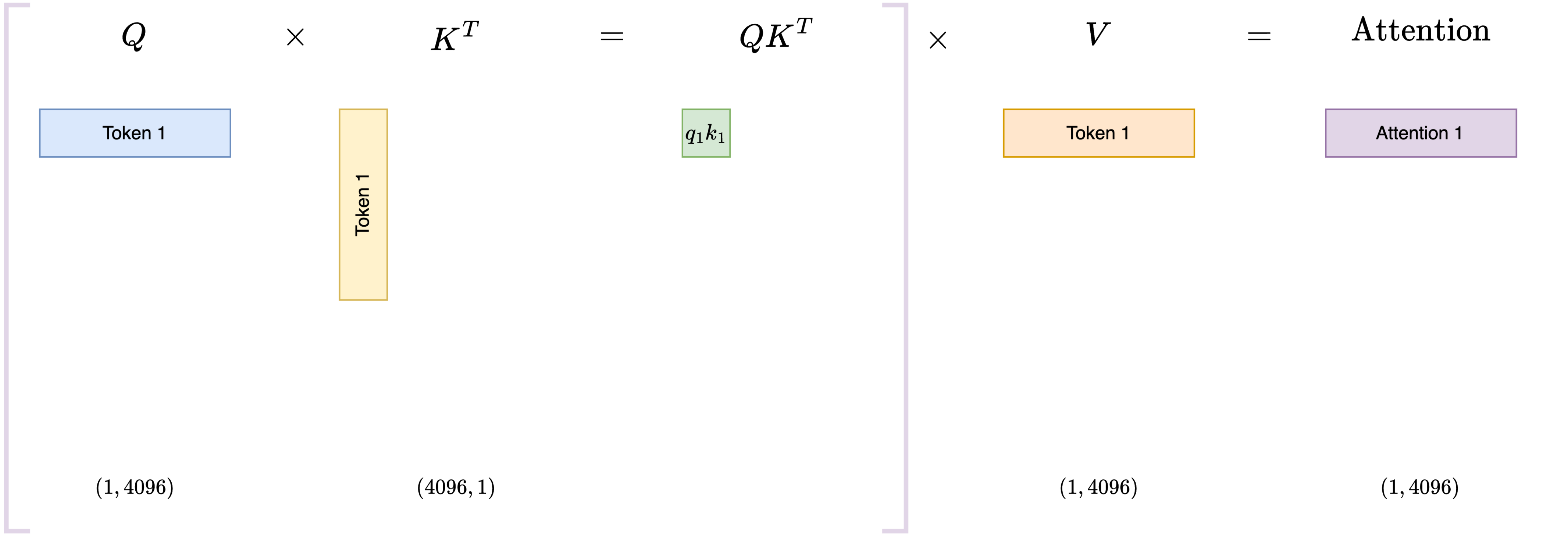
## Inference T=1



$Q \times K^T = QK^T$ $\times\ V = \text{Attention}$

Token 1

Token 1

$q_1 k_1$

Token 1

Attention 1

$(1, 4096)$

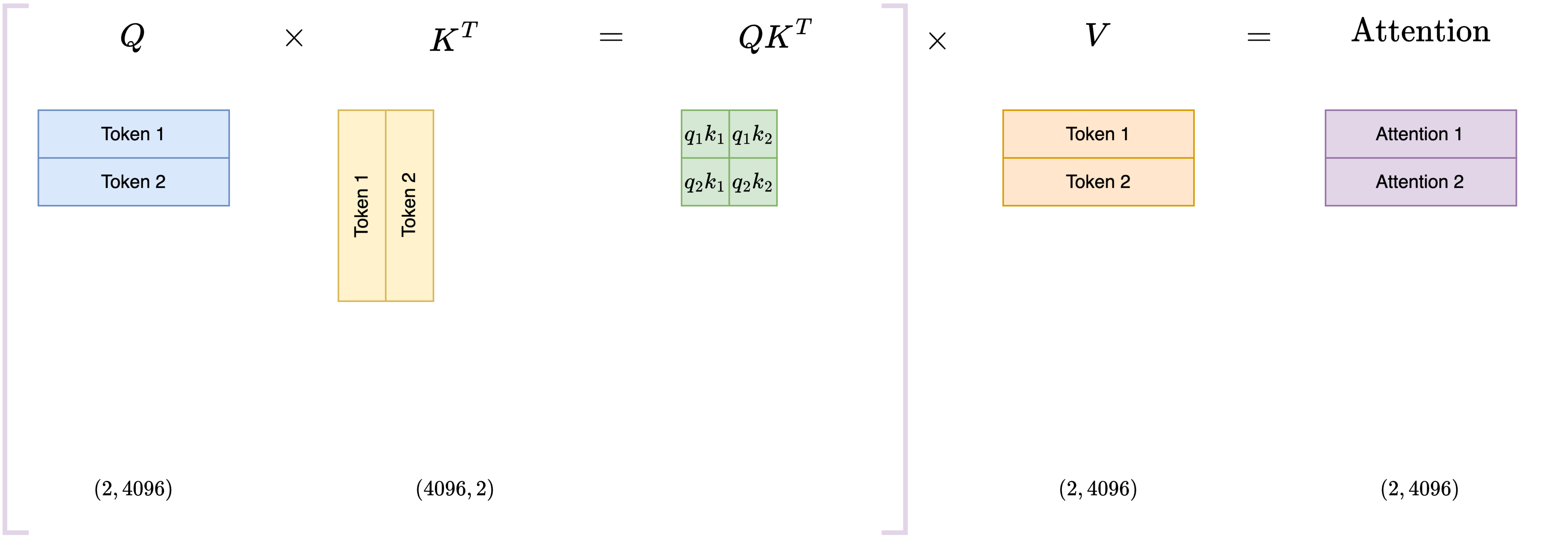
$(4096, 1)$

$(1, 4096)$

$(1, 4096)$

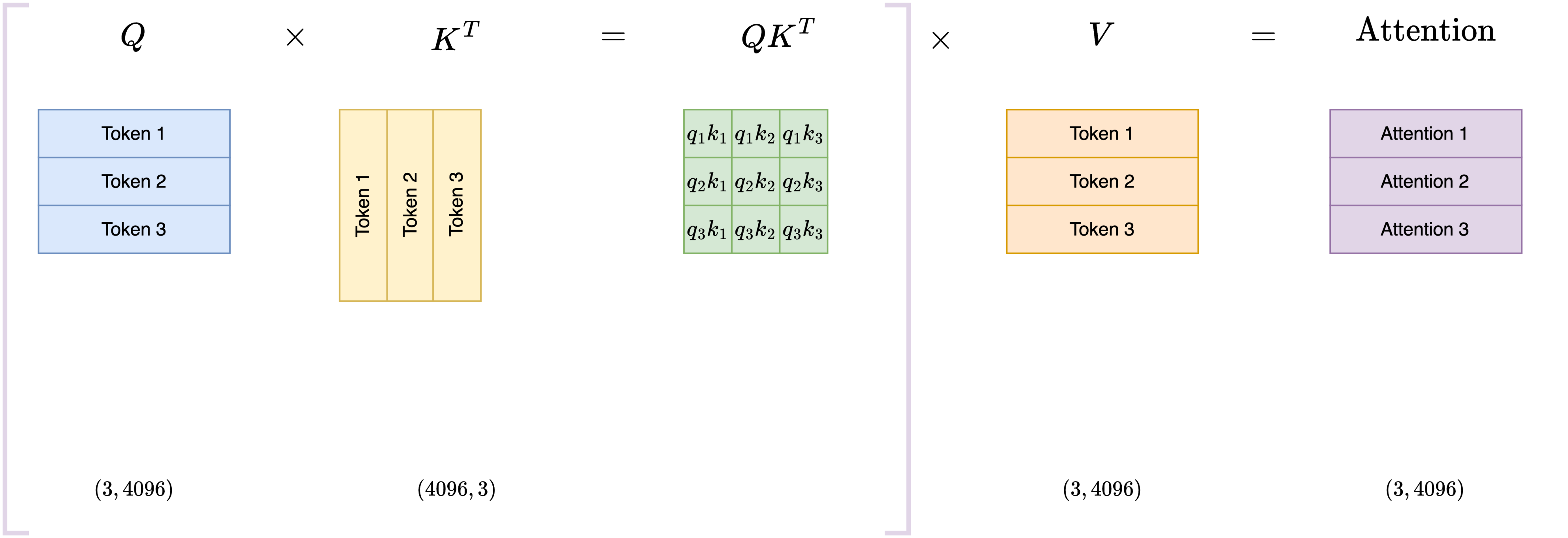
$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

## Inference T=2



$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

## Inference T=3



$Q$ × $K^T$ = $QK^T$ × $V$ = $\text{Attention}$

Token 1

Token 2

Token 3

$(3, 4096)$

Token 1

Token 2

Token 3

$(4096, 3)$

| | | |
|---|---|---|
| $q_1k_1$ | $q_1k_2$ | $q_1k_3$ |
| $q_2k_1$ | $q_2k_2$ | $q_2k_3$ |
| $q_3k_1$ | $q_3k_2$ | $q_3k_3$ |

Token 1

Token 2

Token 3

$(3, 4096)$

Attention 1

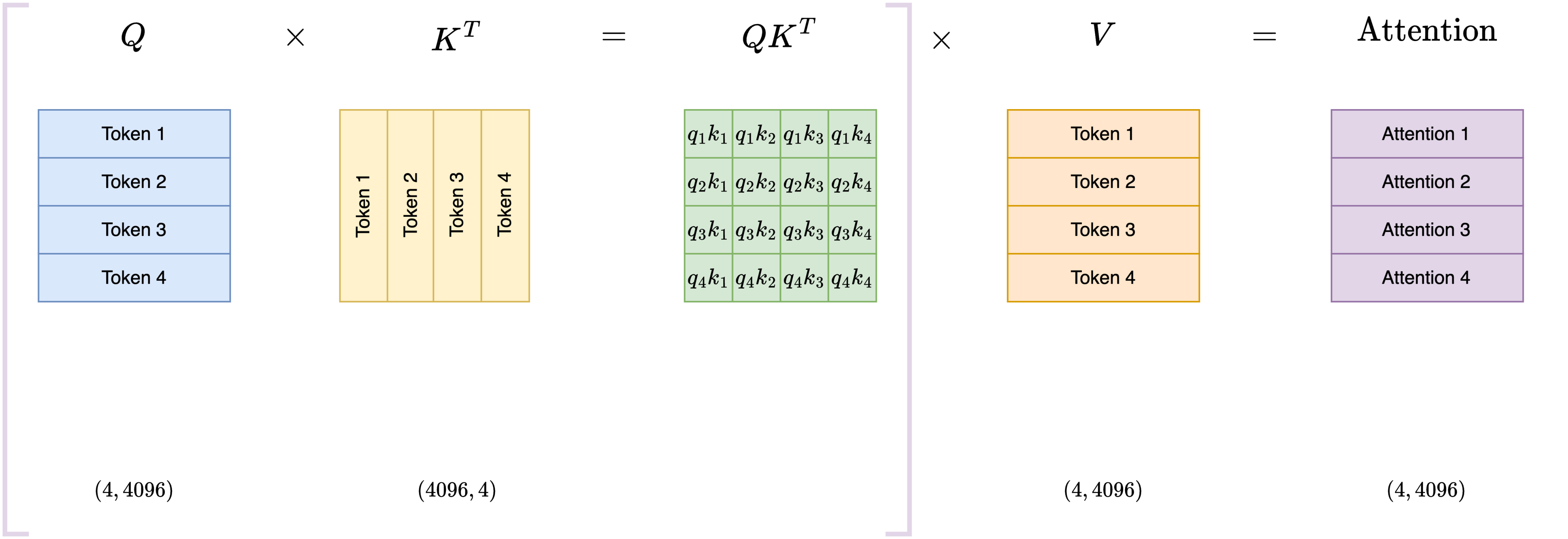
Attention 2

Attention 3

$(3, 4096)$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

## Inference T=4



$$\left[ Q \times K^T = QK^T \right] \times V = \text{Attention}$$

$Q$: Token 1, Token 2, Token 3, Token 4 — $(4, 4096)$

$K^T$: Token 1, Token 2, Token 3, Token 4 — $(4096, 4)$

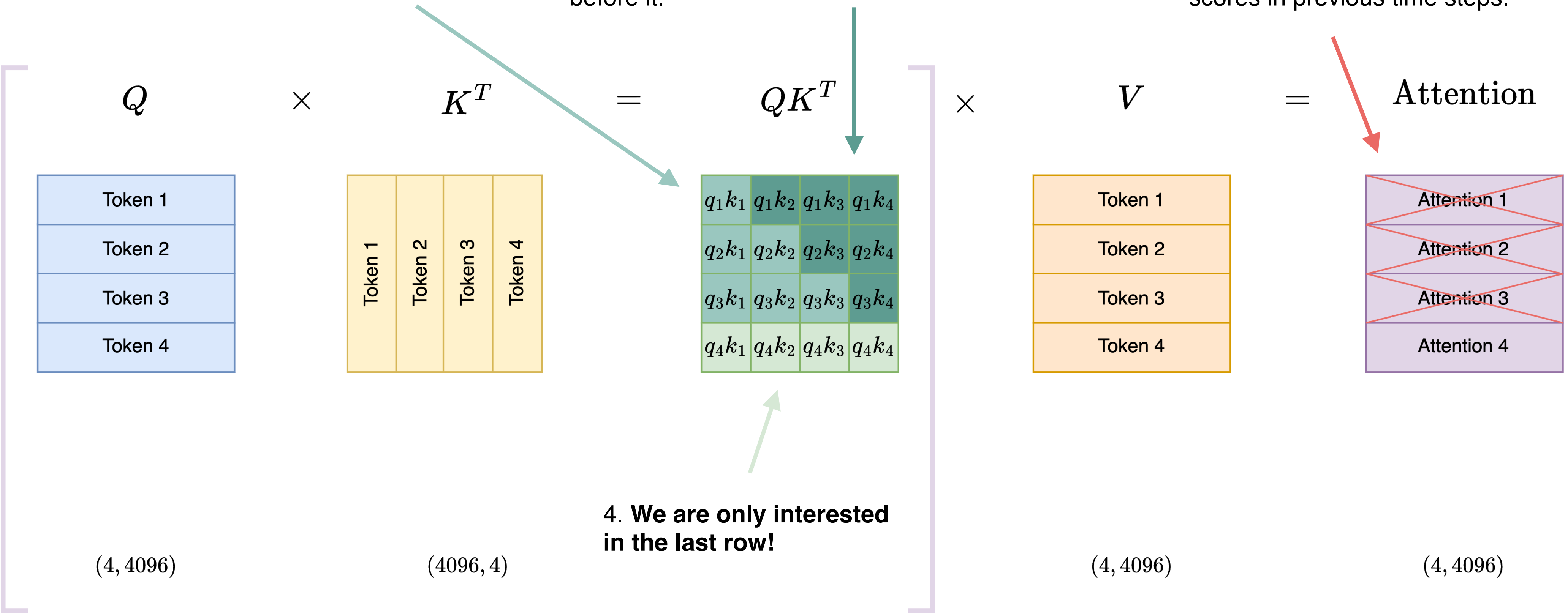
$QK^T$:

| | | | |
|---|---|---|---|
| $q_1k_1$ | $q_1k_2$ | $q_1k_3$ | $q_1k_4$ |
| $q_2k_1$ | $q_2k_2$ | $q_2k_3$ | $q_2k_4$ |
| $q_3k_1$ | $q_3k_2$ | $q_3k_3$ | $q_3k_4$ |
| $q_4k_1$ | $q_4k_2$ | $q_4k_3$ | $q_4k_4$ |

$V$: Token 1, Token 2, Token 3, Token 4 — $(4, 4096)$

Attention: Attention 1, Attention 2, Attention 3, Attention 4 — $(4, 4096)$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$



1. We already computed these dot products in the previous time steps. **Can we cache them?**

2. Since the model is causal, **we don't care about the attention of a token with its successors**, but only with the tokens before it.

3. **We don't care about these,** as we want to predict the next token and we have already predicted these attention scores in previous time steps.

4. **We are only interested in the last row!**

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$