kV-Cache

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_{head}}} + \text{MASK}\right) =$$

	K: I	K: LOVE	K: PEPPERONI
Q: I	1.0	0	0
Q: LOVE	0.6	0.4	0
Q: PEPPERONI	0.2	0.4	0.4

	K: I	K: LOVE	K: PEPPERONI
Q: I	1.0	0	0
Q: LOVE	0.6	0.4	0
Q: PEPPERONI	0.2	0.4	0.4

(3, 3)

$\times$

V

[1 ... 128]	I
[1 ... 128]	LOVE
[1 ... 128]	PEPPERONI

(3, 128)

$=$

[1 ... 128]	I
[1 ... 128]	I LOVE
[1 ... 128]	I LOVE PEPPERONI

(3, 128)

} CONTEXTUALIZED EMBEDDINGS