Masking: We don't want to get affected by future tokens.

	Key1	Key2	Key3	KeyN
Query1	s_{11}	$-\infty$	$\cdots$	$-\infty$
Query2	s_{21}	s_{22}	$\cdots$	$-\infty$
Query3	s_{31}	s_{32}	$\cdots$	$-\infty$
QueryN	s_{N1}	s_{N2}	$\cdots$	s_{NN}

Masked QK^T with upper-triangular $-\infty$

	Key1	Key2	Key3	KeyN
Query1	s_{11}	$-\infty$	$\cdots$	$-\infty$
Query2	s_{21}	α_{22}	$\cdots$	$-\infty$
Query3	$\vdots$	$\vdots$	$\ddots$	$\vdots$
QueryN	s_{N1}	s_{N2}	$\cdots$	s_{NN}

Softmax applied per query row (row highlighted)

	Key1	Key2	Key3	KeyN
Query1	α_{11}	0	$\cdots$	0
Query2	α_{21}	α_{22}	$\cdots$	0
Query3	α_{31}	α_{32}	$\cdots$	0
QueryN	α_{N1}	α_{N2}	$\cdots$	α_{NN}

Attention Weight(After softmax): masked positions ($-\infty$) become 0