ce 1	rmsnorm att	
Device 1	q	
Device 2	k	Direction
Devi	V	ion of
	multihead att	Inference
:	att	ence.
Device n	rmsnorm ffn	
Devi	ffn	$\downarrow$