$$Q = xW_q + b_q$$
$$K = xW_k + b_k$$
$$V = xW_v + b_v$$

BMM

$$Q = \text{transpose}(Q)$$
$$K = \text{transpose}(K)$$
$$V = \text{transpose}(V)$$
$$O = \text{softmax}(QK^T)V$$

Flash Attention

$$O = \text{transpose}(O)$$
$$O = OW_o + b$$
$$\cdots\cdots$$

Fused with Layer Norm