Scaled Dot-Product Attention MatMul SoftMax Mask (opt.) Scale MatMul