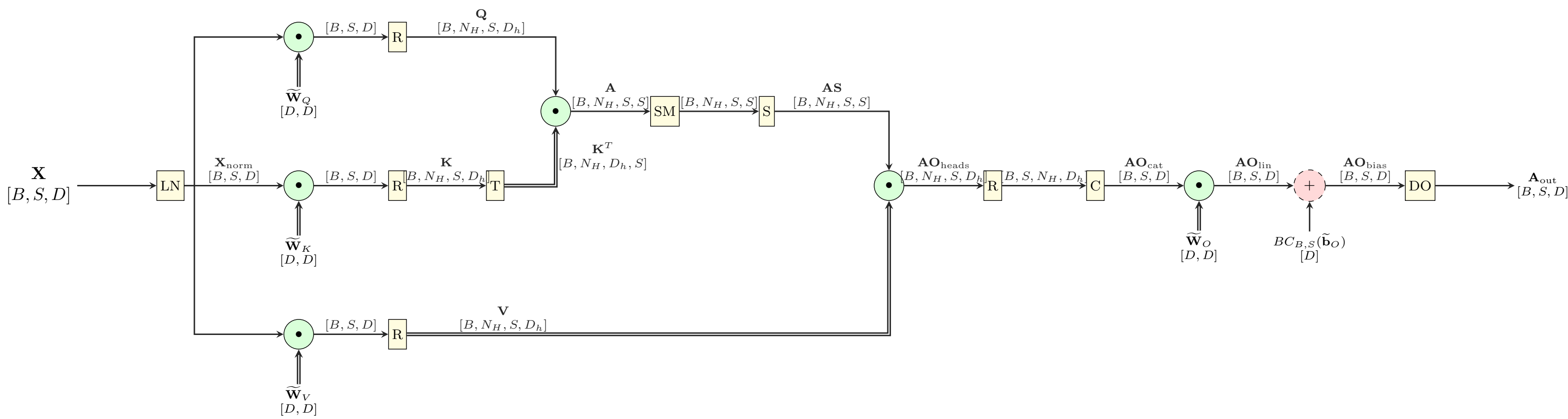
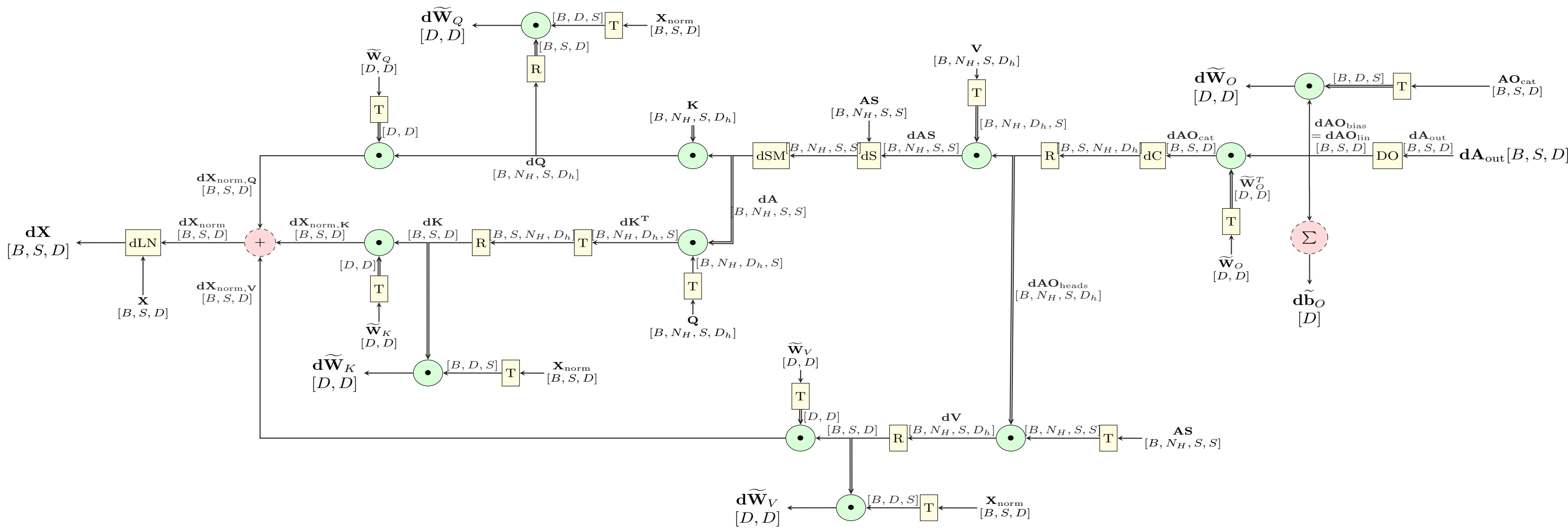
Multi-Head Attention Forward Pass



Multi-Head Attention Backward Pass



	Abbrev	Name	Type / Shape	Notes
Operations (Ops)	LN	Layer Normalization	op	Normalizes per token (per last dim D).
	DO	Dropout	op	Training-time stochastic dropout.
	+	Bias Add	op	Adds broadcast bias; see $\mathrm{BC}_{B,S}(\cdot)$.
	T	Transpose	op	Context-dependent dims (e.g., $[B, N_H, S, D_h] \to [B, N_H, D_h, S]$).
	R	Reshape / Split / Merge	op	Head split/merge: $[B, S, D] \leftrightarrow [B, N_H, S, D_h]$.
	C	Concatenate	op	Join heads along last dim: $[B, S, N_H, D_h] \to [B, S, D]$.
	SM	Scale (+ Mask)	op	Multiply by $1/\sqrt{D_h}$ and apply mask to scores.
	S	Softmax	op	Softmax over key length (S) per head.
	$\mathrm{BC}_{B,S}(\cdot)$	Broadcast	op	Broadcast a length-D (or D_h, V) bias to $[B, S, \cdot]$.
	dS	Softmax Backward	op	Backprop through softmax over S.
	dSM	Scale/Mask Backward	op	Backprop through scaling and masking.
	dC	De-concatenate (Backward)	op	Split grads from concatenated heads.
	dLN	LayerNorm Backward	op	Uses cached LN stats (μ, σ) and $\mathbf{X}$.

Data Tensors (Values)

Symbol	Name	Shape	Notes
$\mathbf{X}$	Input hidden states	$[B, S, D]$	Into MHA block (pre-LN).
$\mathbf{X}_{\text{norm}}$	LN output	$[B, S, D]$	Result of LN($\mathbf{X}$).
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query/Key/Value	$[B, N_H, S, D_h]$	From linear projections of $\mathbf{X}_{\text{norm}}$.
$\widetilde{\mathbf{W}}_Q$	Q weight	$[D, D]$	Per-head realized by reshape; drawn as single matmul.
$\widetilde{\mathbf{W}}_K$	K weight	$[D, D]$	Same convention.
$\widetilde{\mathbf{W}}_V$	V weight	$[D, D]$	Same convention.
$\widetilde{\mathbf{W}}_O$	Output-proj weight	$[D, D]$	Maps concatenated heads back to model dim.
$\widetilde{\mathbf{b}}_O$	Output bias	$[D]$	Broadcast via $\text{BC}_{B,S}$.
$\mathbf{A}$	Attention scores	$[B, N_H, S, S]$	$\mathbf{QK}^T/\sqrt{D_h}$ (plus mask).
$\mathbf{AS}$	Attention weights	$[B, N_H, S, S]$	softmax($\mathbf{A}$).
$\mathbf{AO}_{\text{heads}}$	Per-head outputs	$[B, N_H, S, D_h]$	$\mathbf{AS} \cdot \mathbf{V}$.
$\mathbf{AO}_{\text{cat}}$	Concatenated heads	$[B, S, D]$	After C.
$\mathbf{AO}_{\text{lin}}$	Linear output	$[B, S, D]$	$\mathbf{AO}_{\text{cat}}\widetilde{\mathbf{W}}_O$.
$\mathbf{AO}_{\text{bias}}$	Bias-added output	$[B, S, D]$	$\mathbf{AO}_{\text{lin}} + \widetilde{\mathbf{b}}_O$.
$\mathbf{A}_{\text{out}}$	MHA output	$[B, S, D]$	After dropout; to next sublayer.
$\mathbf{dA}_{\text{out}}$	Grad wrt MHA output	$[B, S, D]$	Backprop signal entering MHA.
$\mathbf{dQ}, \mathbf{dK}, \mathbf{dV}$	Gradients for Q/K/V	$[B, N_H, S, D_h]$	From attention-core backward.
$\mathbf{dK^T}$	Grad of K^T	$[B, N_H, D_h, S]$	Before transpose/reshape to $\mathbf{dK}$.
$\mathbf{dAO}_{\text{heads}}$	Grad at heads	$[B, N_H, S, D_h]$	Split from $\mathbf{dAO}_{\text{cat}}$.
$\mathbf{dX}_{\text{norm},Q}$	Grad wrt X_{norm} (Q branch)	$[B, S, D]$	Contribution via W_Q^T.
$\mathbf{dX}_{\text{norm},K}$	Grad wrt X_{norm} (K branch)	$[B, S, D]$	Contribution via W_K^T.
$\mathbf{dX}_{\text{norm},V}$	Grad wrt X_{norm} (V branch)	$[B, S, D]$	Contribution via W_V^T.
$\mathbf{dX}_{\text{norm}}$	Sum of above	$[B, S, D]$	Input to dLN.
$\mathbf{dX}$	Grad wrt input X	$[B, S, D]$	Output of dLN.
$\mathbf{d}\widetilde{\mathbf{W}}_Q$	Q weight grad	$[D, D]$	Standard matmul rule.
$\mathbf{d}\widetilde{\mathbf{W}}_K$	K weight grad	$[D, D]$	Standard matmul rule.
$\mathbf{d}\widetilde{\mathbf{W}}_V$	V weight grad	$[D, D]$	From reshaped $d\mathbf{V}$ and X_{norm}.
$\mathbf{d}\widetilde{\mathbf{W}}_O$	Output-proj grad	$[D, D]$	From $\mathbf{AO}_{\text{cat}}^T$ and $d\mathbf{AO}_{\text{lin}}$.
$\mathbf{d}\widetilde{\mathbf{b}}_O$	Output bias grad	$[D]$	Sum over B, S of $d\mathbf{AO}_{\text{lin}}$.

Shape symbols: B=batch size, S=sequence length, D=model dim, N_H=num heads, $D_h = D/N_H$.

Implementation note: Per-head $[D, D]$ drawings depict fused linears realized via reshape to $N_H \times D_h$.