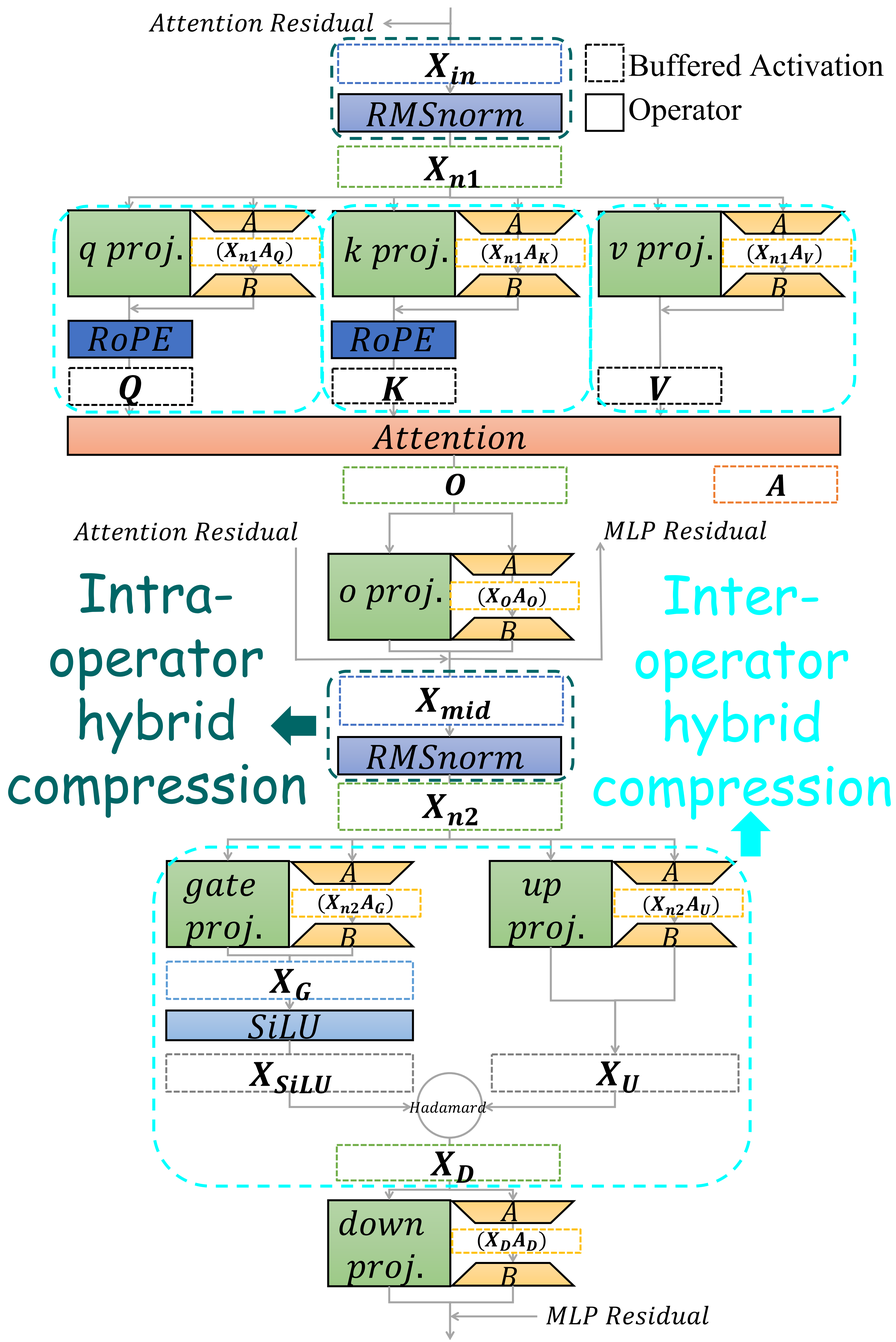

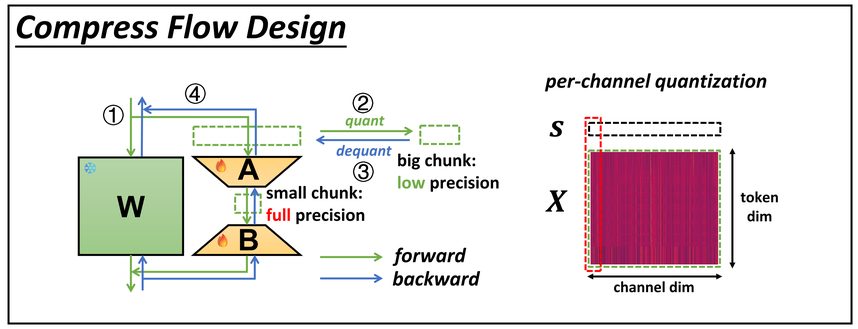

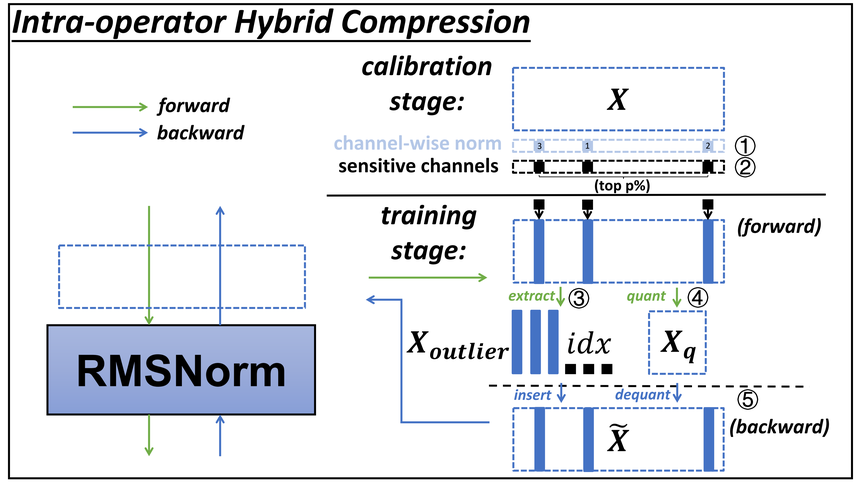

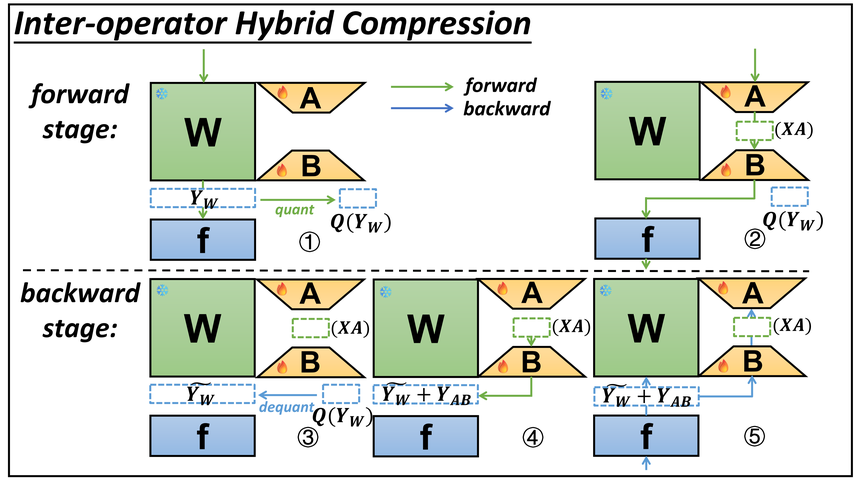

Operators	Buffered Activation	Memory Usage
$X_{n1} = RMSNorm(X_{in})$	X_{in} σ^2_{in}	(b,s,d) (b,s)
$Q = X_{n1}(W_Q + A_QB_Q)$ $K = X_{n1}(W_K + A_KB_K)$ $V = X_{n1}(W_V + A_VB_V)$	X_{n1} $(X_{n1}A_Q), (X_{n1}A_K)$ $(X_{n1}A_V)$	(b,s,d) $3 \times (b,s,r)$
$Q = RoPE(Q, cos, sin)$ $K = RoPE(K, cos, sin)$	cos sin	$2 \times (s, d/h)$
$S = QK^T, A = Softmax(S)$ $O = AV$ w/o FlashAttn	Q,K,V A	$3 \times (b,s,d)$ (b,h,s,s)
$O = FlashAttn(Q,K,V)$ w FlashAttn	Q,K,V	$3 \times (b,s,d)$
$X_{mid} = O(W_O + A_OB_O)$	O (OA_O)	(b,s,d) (b,s,r)
$X_{n2} = RMSNorm(X_{mid})$	X_{mid} σ^2_{mid}	(b,s,d) (b,s)
$X_G = X_{n2}(W_G + A_GB_G)$ $X_U = X_{n2}(W_U + A_UB_U)$	X_{n2} $(X_{n2}A_G), (X_nA_U)$	(b,s,d) $2 \times (b,s,r)$
$X_{SiLU} = SiLU(X_G)$	X_G	(b,s,d_f)
$X_D = X_{SiLU} \odot X_U$	X_{SiLU} X_U	(b,s,d_f) (b,s,d_f)
$X_{out} = X_D(W_D + A_DB_D)$	X_D (X_DA_D)	(b,s,d_f) (b,s,r)
$Estimated\ Total\ Size(bit)$ $+HyCLoRA@raw\ quant$ $+HyCLoRA@inter + intra$		$(8d + 4d_f)bsw$ $(8d + 4d_f)bsw_q$ $(8d + 2d_f)bsw_q$