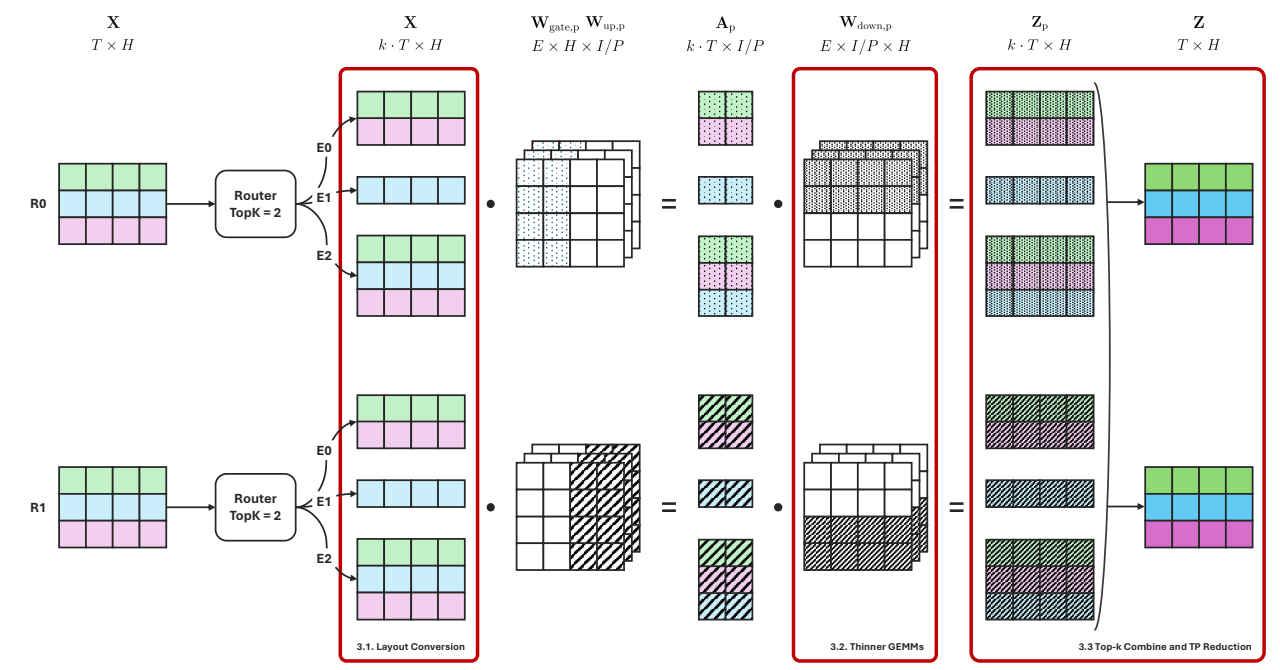

X
$T \times H$
X
$k \cdot T \times H$
$\mathbf{W}_{\text{gate,p}}$ $\mathbf{W}_{\text{up,p}}$
$E \times H \times I/P$
$\mathbf{A}_{\text{p}}$
$k \cdot T \times I/P$
$\mathbf{W}_{\text{down,p}}$
$E \times I/P \times H$
$\mathbf{Z}_{\text{p}}$
$k \cdot T \times H$
Z
$T \times H$
R0
R1
Router
TopK = 2
E0
E1
E2
3.1. Layout Conversion
3.2. Thinner GEMMs
3.3 Top-k Combine and TP Reduction