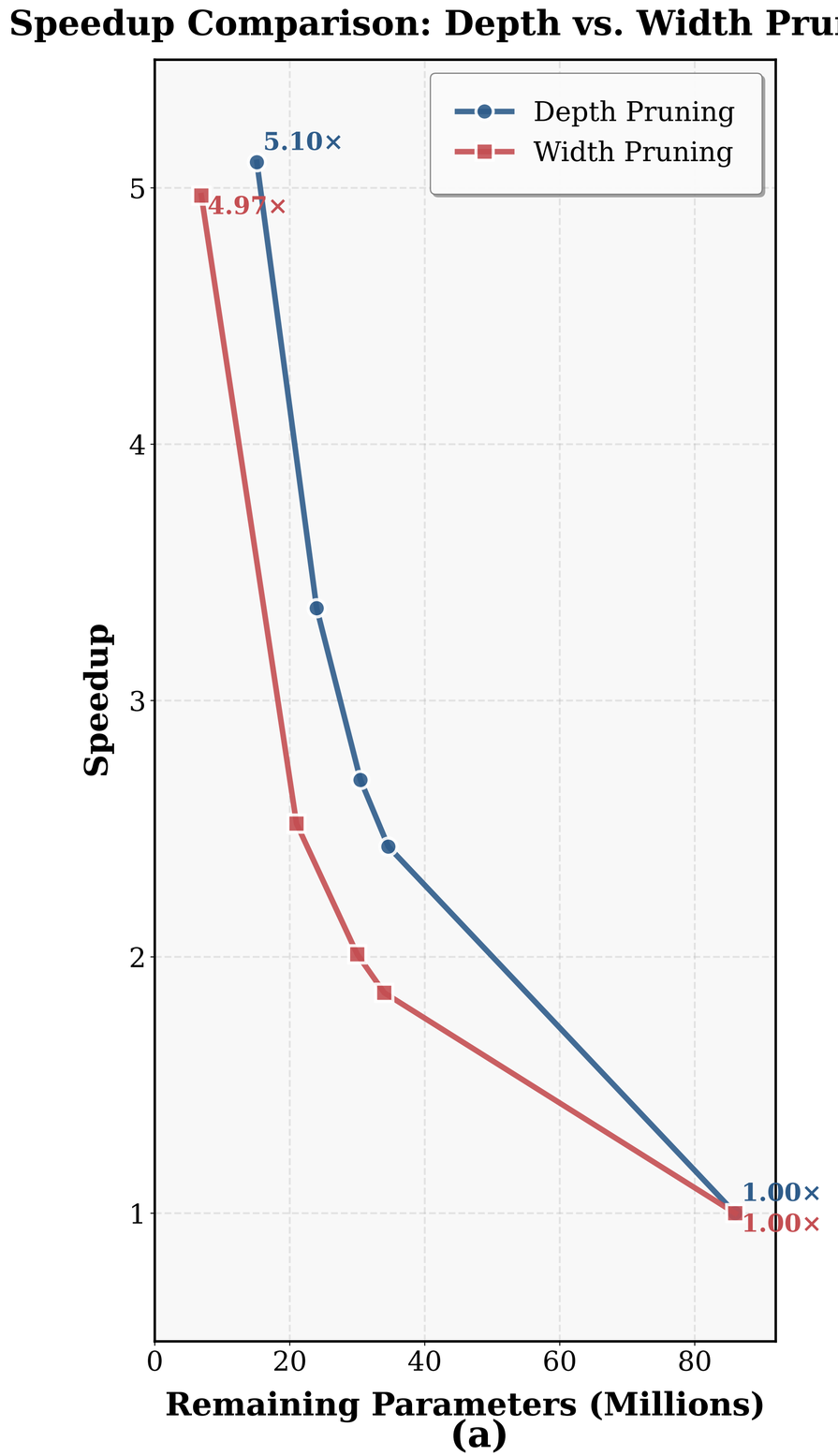
Speedup Comparison: Depth vs. Width Pruning
Depth Pruning
Width Pruning
5.10×
4.97×
1.00×
1.00×
Speedup
Remaining Parameters (Millions)
0
20
40
60
80
1
2
3
4
5

(a)

Computation Cost Breakdown of ViT Components
DeiT-Base 224×224
49.4%
43.1%
Activation 1.19%
LayerNorm 3.37%
Other 2.96%
DeiT-Base 384×384
51.9%
41.7%
Activation 1.22%
LayerNorm 4.24%
Other 0.87%
Components
Attention
Linear
Activation
LayerNorm
Other
Swin-Large 224×224
58.3%
24.9%
11.2%
Activation 2.55%
Other 3.05%
Swin-Base 224×224
57.0%
27.1%
12.5%
Activation 1.61%
Other 1.80%

(b)