Performance for Model: 1B, Seq Len: 1024 on RTX5090 176 137.00 168.44 168.33 168.50 168.37 160 168.61 168.44 168.29 168.20 136.88 - 160 144 136.39 168.27 167.93 168.24 168.32 128 -140 136.02 168.27 168.18 168.43 168.33 Host Memory (GB) 136.74 168.26 167.84 168.06 168.22 120 136.13 168.43 168.15 168.21 168.14 136.06 168.26 168.09 167.98 167.93 100 64 137.36 168.13 168.11 167.92 167.96 48 137.66 168.21 167.95 167.90 168.06 80 32 168.20 167.71 137.23 167.63 167.71 60 16 167.68 136.38 167.61 167.86 167.59 16 20 28 24 30 Device Memory (GB)