Performance for Model: 1B, Seq Len: 4096 on RTX5090 176 140.48 159.77 159.89 159.70 159.67 160 140.45 160.07 159.79 159.70 159.89 - 160 144 139.62 159.74 159.89 159.84 160.00 128 - 140 139.60 159.69 159.60 159.65 159.61 Host Memory (GB) 140.50 159.76 159.73 159.64 159.87 120 140.00 159.69 159.54 159.63 159.74 139.18 159.55 159.50 159.72 159.72 - 100 64 140.95 159.71 159.64 159.68 159.57 48 139.85 159.57 159.39 159.63 159.79 80 32 140.21 159.42 159.10 159.56 159.44 - 60 16 140.10 159.20 159.05 158.93 159.31 16 20 28 24 30 Device Memory (GB)