1) As our Tensor Casting primarily resolves bottlenecks incurred by gradient expand-coalesce, this graph is intended to visualize how much faster gradient expand-coalesce becomes when applied to Ours(CPU) and Ours(NMP), respectively. Results are normalized to Baseline(CPU) at 1.0 (i.e., higher the better)