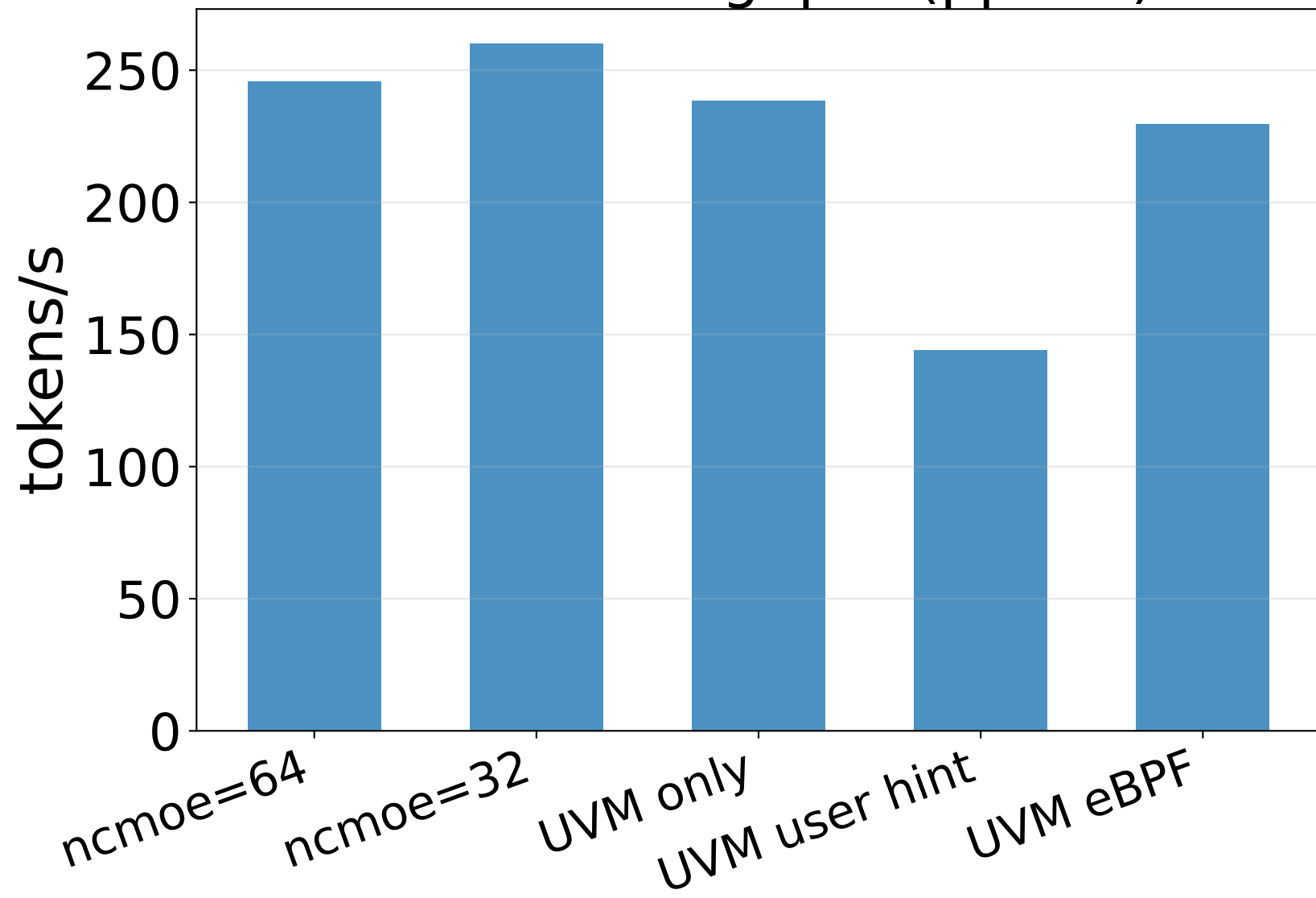**Prefill Throughput (pp512)** and **Decode Throughput (tg128)**

Categories (x-axis): ncmoe=64, ncmoe=32, UVM only, UVM user hint, UVM eBPF

Y-axis: tokens/s