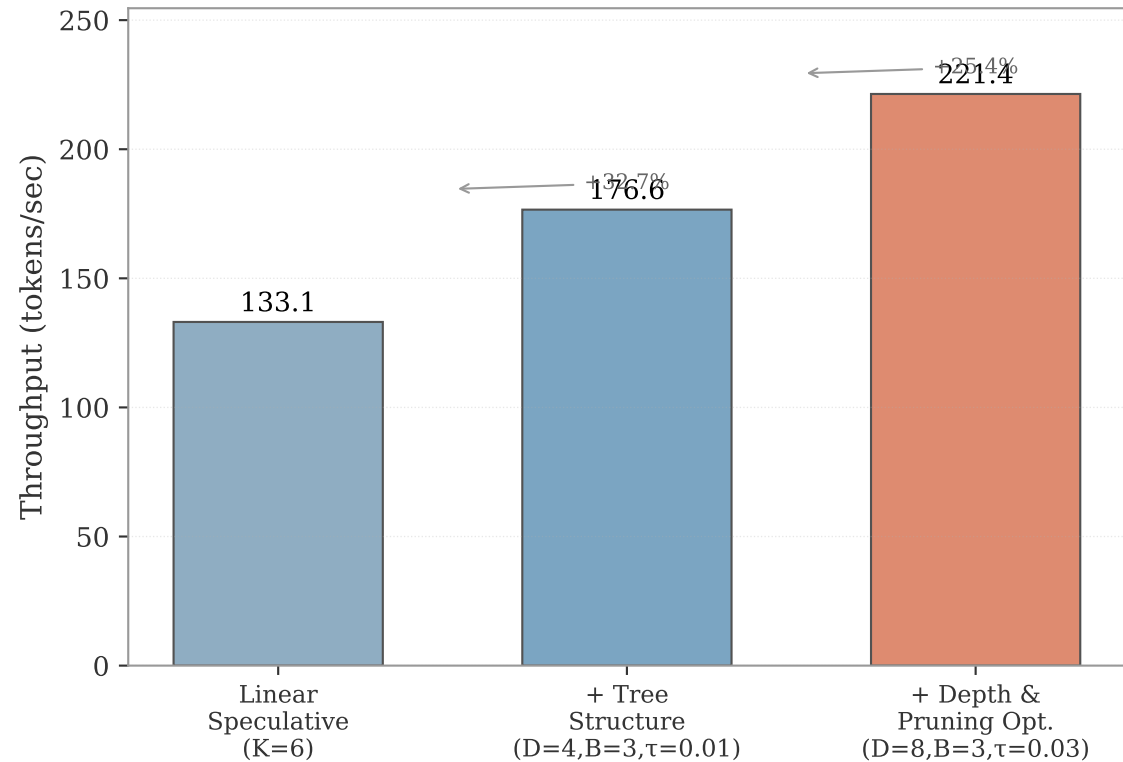**(a) Throughput Improvement**

(b) Speedup Improvement

Figure (a) — Throughput (tokens/sec):
- Linear Speculative (K=6): 133.1
- + Tree Structure (D=4,B=3,τ=0.01): 176.6 (+32.7%)
- + Depth & Pruning Opt. (D=8,B=3,τ=0.03): 221.4 (+25.4%)

Figure (b) — Speedup (vs AR baseline):
- Linear Speculative (K=6): 1.11×
- + Tree Structure (D=4,B=3,τ=0.01): 1.43× (+0.32×)
- + Depth & Pruning Opt. (D=8,B=3,τ=0.03): 1.79× (+0.36×)