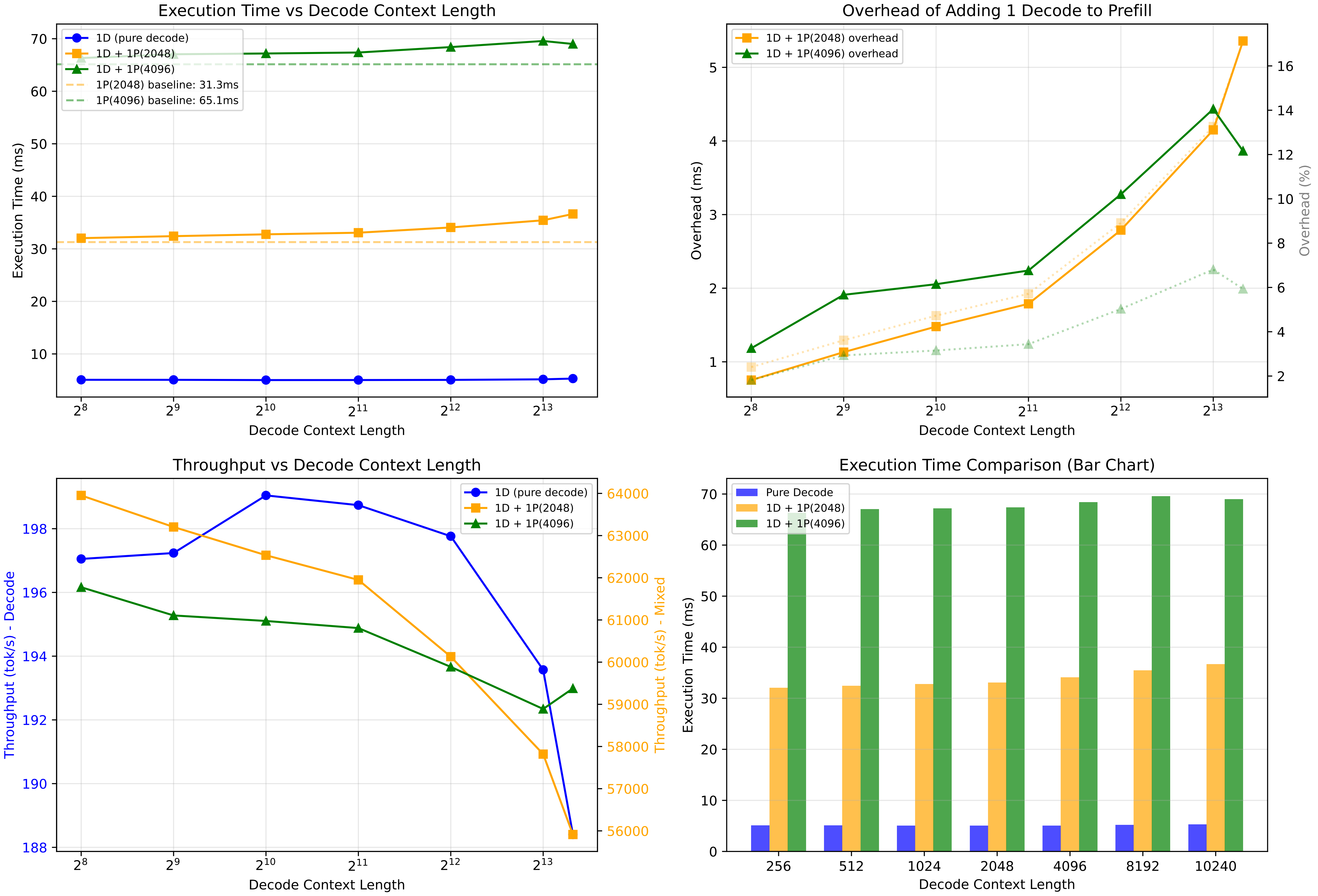
vLLM Benchmark Results - Qwen/Qwen3-4B


Execution Time vs Decode Context Length
Execution Time (ms)
1D (pure decode)
1D + 1P(2048)
1D + 1P(4096)
1P(2048) baseline: 31.3ms
1P(4096) baseline: 65.1ms
2^{8}
2^{9}
2^{10}
2^{11}
2^{12}
2^{13}
Decode Context Length
Overhead of Adding 1 Decode to Prefill
1D + 1P(2048) overhead
1D + 1P(4096) overhead
Overhead (ms)
Overhead (%)
2^{8}
2^{9}
2^{10}
2^{11}
2^{12}
2^{13}
Decode Context Length
Throughput vs Decode Context Length
1D (pure decode)
1D + 1P(2048)
1D + 1P(4096)
Throughput (tok/s) - Decode
Throughput (tok/s) - Mixed
2^{8}
2^{9}
2^{10}
2^{11}
2^{12}
2^{13}
Decode Context Length
Execution Time Comparison (Bar Chart)
Pure Decode
1D + 1P(2048)
1D + 1P(4096)
Execution Time (ms)
Decode Context Length