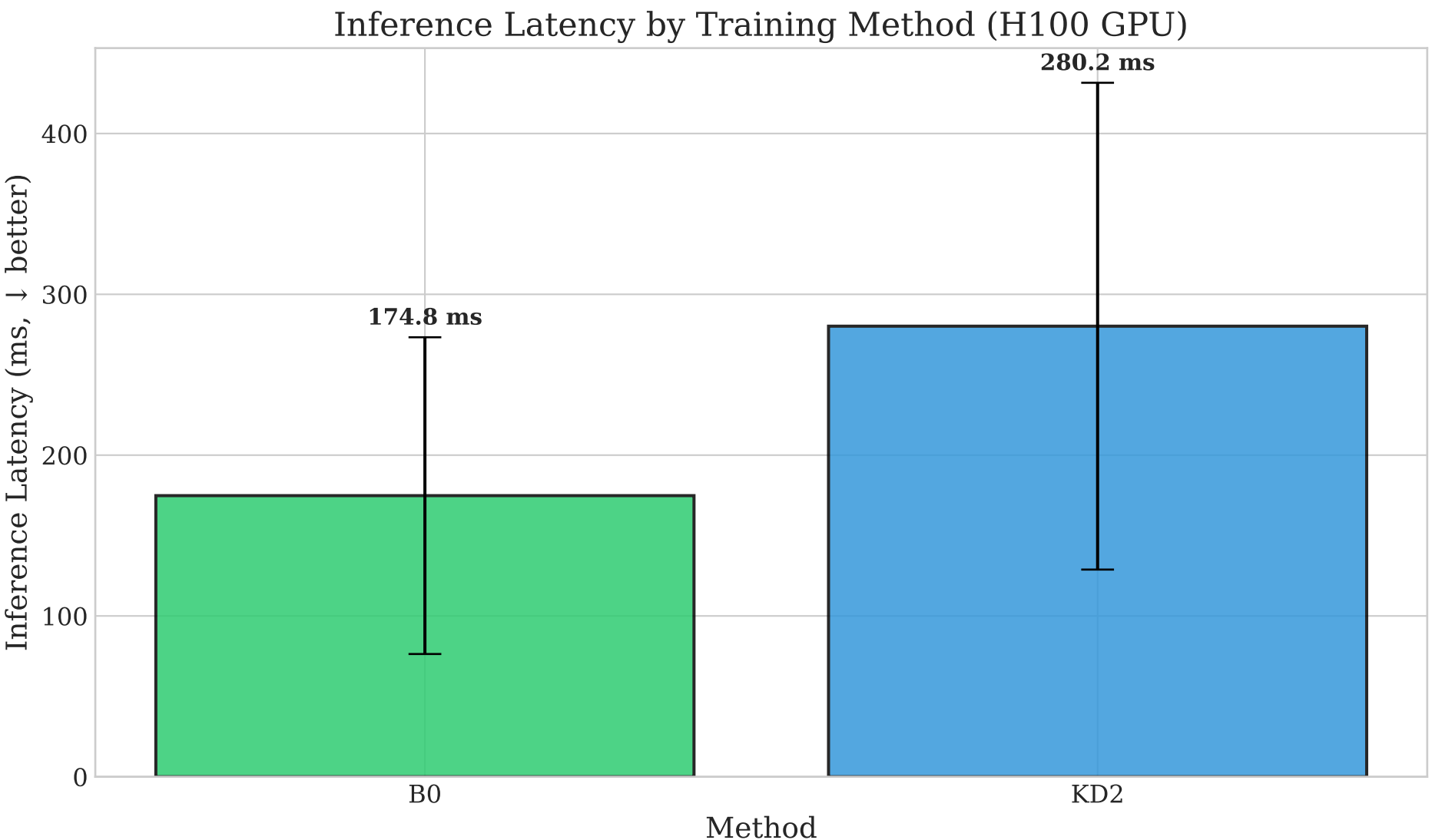
Inference Latency by Training Method (H100 GPU)
Inference Latency (ms, ↓ better)
0
100
200
300
400
174.8 ms
280.2 ms
B0
KD2
Method