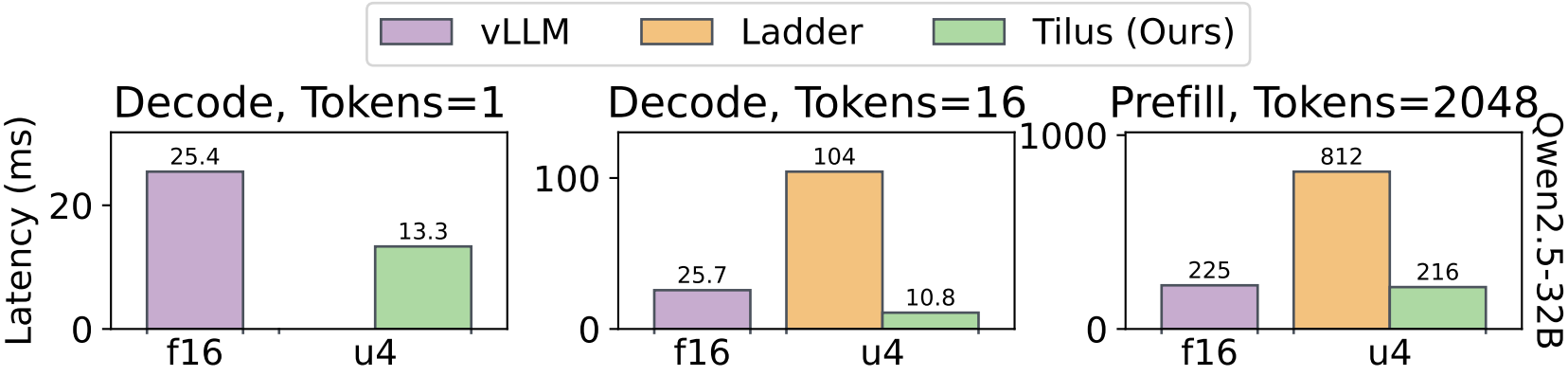
vLLM
Ladder
Tilus (Ours)
Decode, Tokens=1
Decode, Tokens=16
Prefill, Tokens=2048
Latency (ms)
Qwen2.5-32B
0
20
100
1000
25.4
13.3
25.7
104
10.8
225
812
216
f16
u4