

D Throughput (T/S)

vLLM

1

2

4

8

16

Batch Size

120

100

80

60

40

20

1

2

4

8

16

1

2

4

8

16

1

2

4

8

16