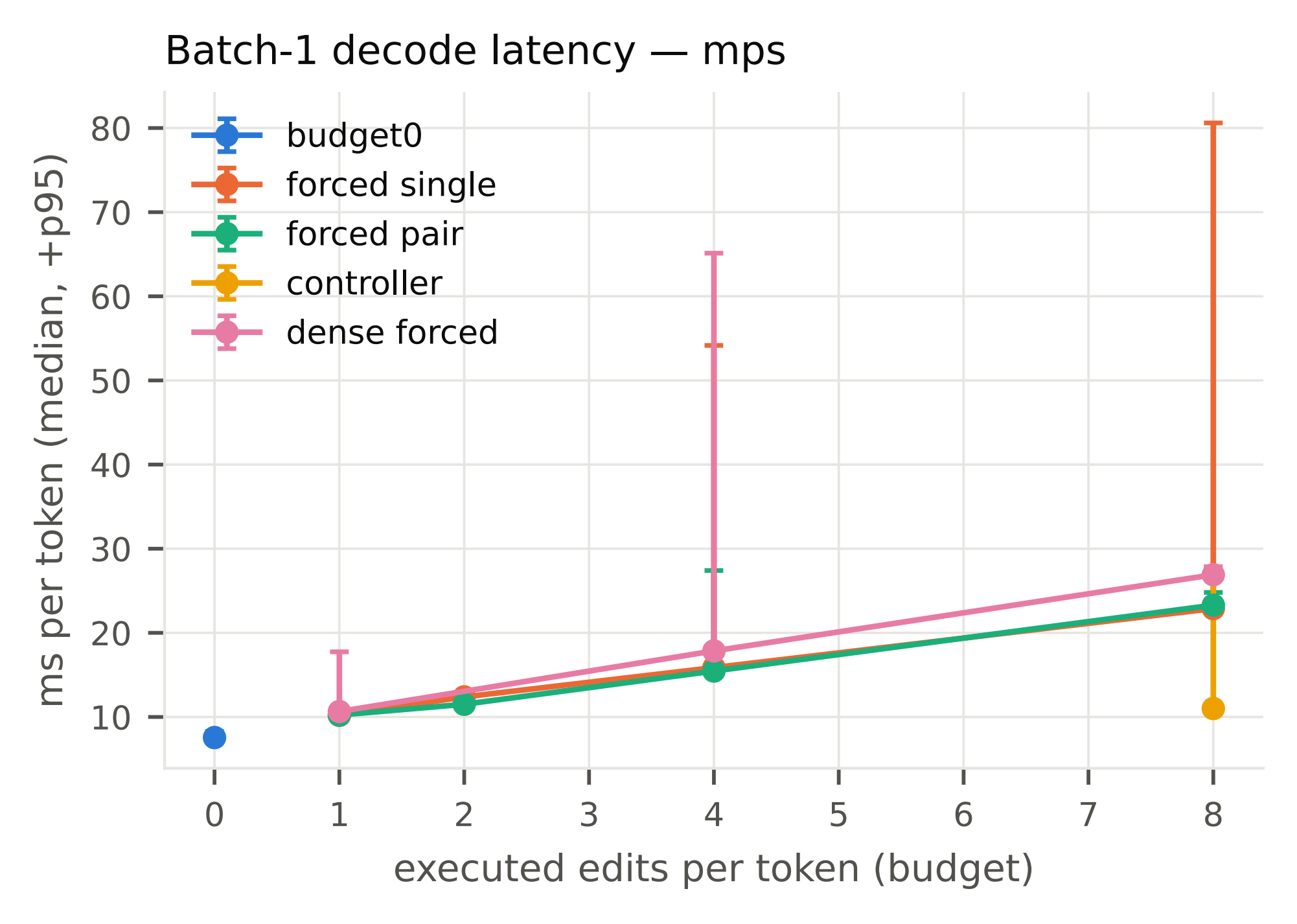

Batch-1 decode latency — mps
budget0
forced single
forced pair
controller
dense forced
ms per token (median, +p95)
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
executed edits per token (budget)