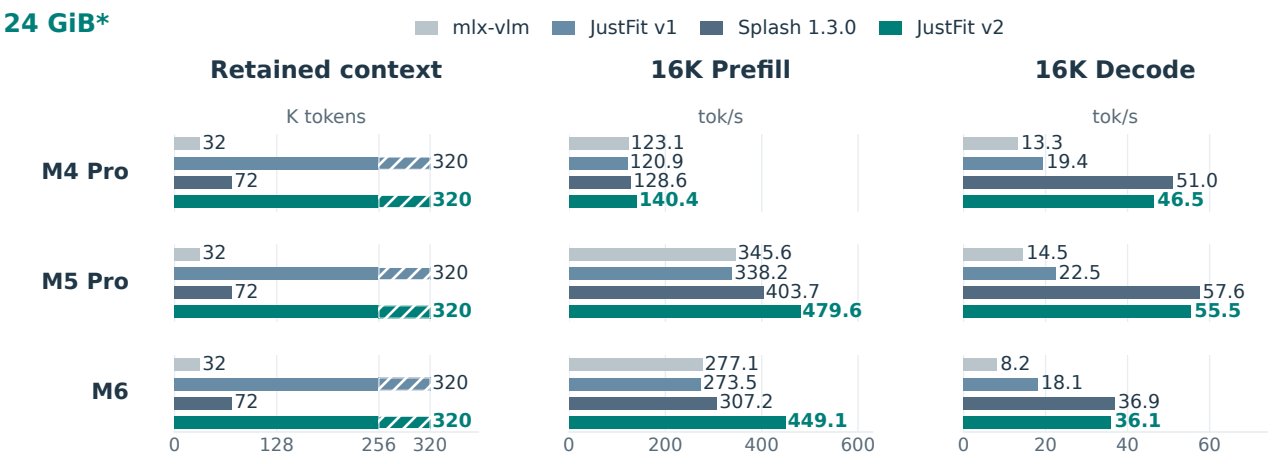

24 GiB*
mlx-vlm
JustFit v1
Splash 1.3.0
JustFit v2
Retained context
16K Prefill
16K Decode
K tokens
tok/s
tok/s
M4 Pro
M5 Pro
M6
32
320
72
320
123.1
120.9
128.6
140.4
13.3
19.4
51.0
46.5
32
320
72
320
345.6
338.2
403.7
479.6
14.5
22.5
57.6
55.5
32
320
72
320
277.1
273.5
307.2
449.1
8.2
18.1
36.9
36.1
0
128
256
320
0
200
400
600
0
20
40
60