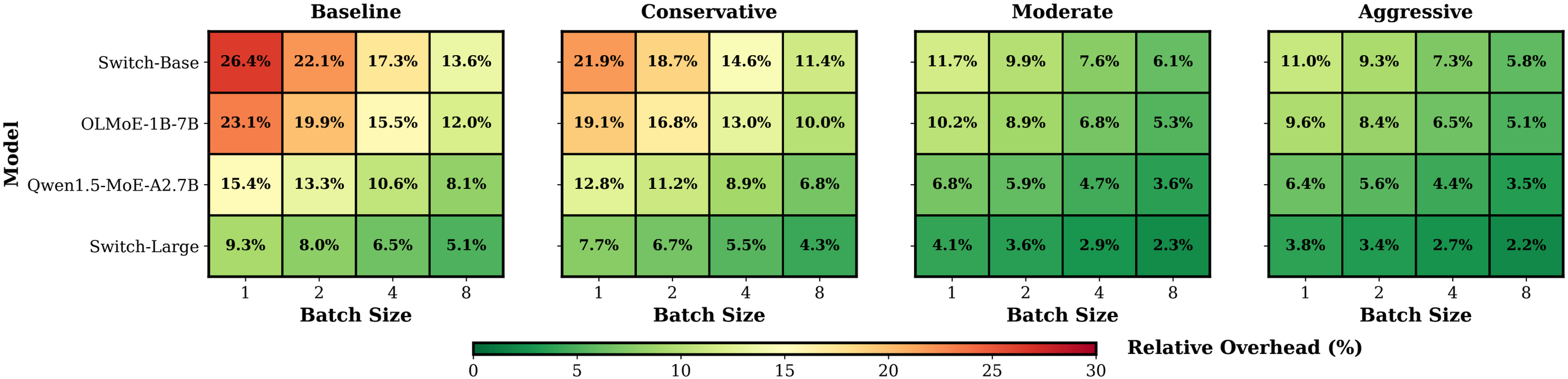
Baseline
Conservative
Moderate
Aggressive
Model
Switch-Base
OLMoE-1B-7B
Qwen1.5-MoE-A2.7B
Switch-Large
26.4% 22.1% 17.3% 13.6%
23.1% 19.9% 15.5% 12.0%
15.4% 13.3% 10.6% 8.1%
9.3% 8.0% 6.5% 5.1%
21.9% 18.7% 14.6% 11.4%
19.1% 16.8% 13.0% 10.0%
12.8% 11.2% 8.9% 6.8%
7.7% 6.7% 5.5% 4.3%
11.7% 9.9% 7.6% 6.1%
10.2% 8.9% 6.8% 5.3%
6.8% 5.9% 4.7% 3.6%
4.1% 3.6% 2.9% 2.3%
11.0% 9.3% 7.3% 5.8%
9.6% 8.4% 6.5% 5.1%
6.4% 5.6% 4.4% 3.5%
3.8% 3.4% 2.7% 2.2%
1 2 4 8
Batch Size
0 5 10 15 20 25 30
Relative Overhead (%)