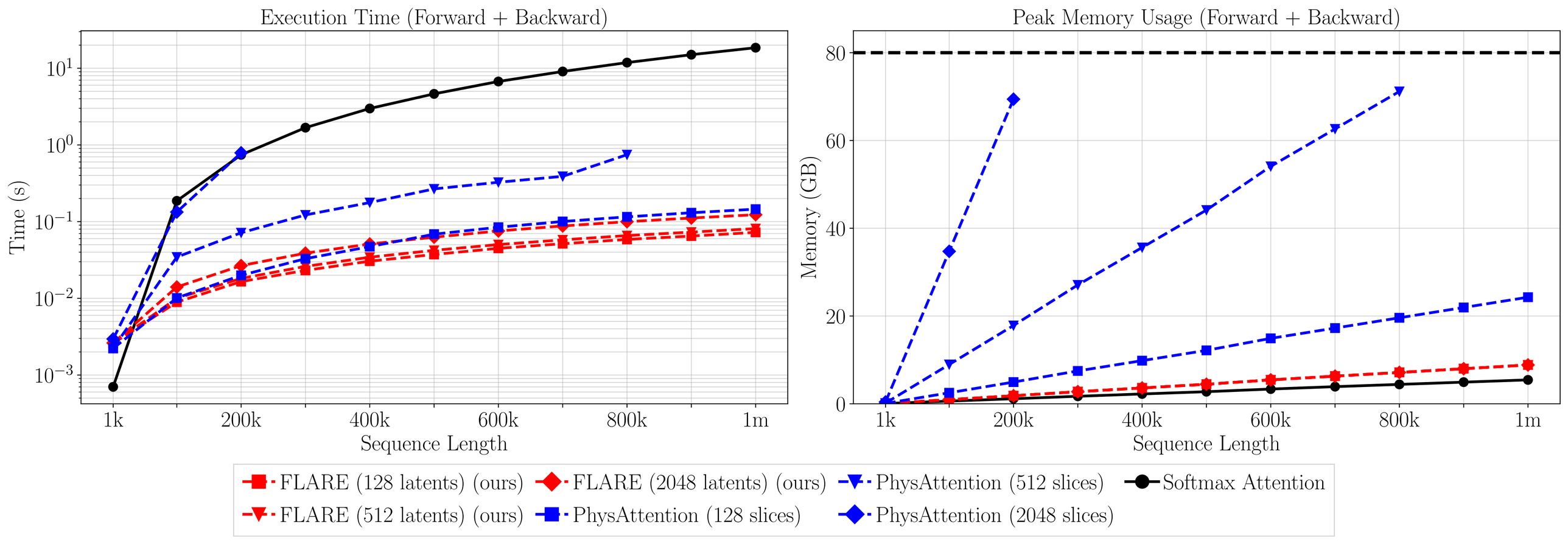
Execution Time (Forward + Backward)
Time (s)
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
1k
200k
400k
600k
800k
1m
Sequence Length
Peak Memory Usage (Forward + Backward)
Memory (GB)
80
60
40
20
0
1k
200k
400k
600k
800k
1m
Sequence Length
FLARE (128 latents) (ours)
FLARE (2048 latents) (ours)
PhysAttention (512 slices)
Softmax Attention
FLARE (512 latents) (ours)
PhysAttention (128 slices)
PhysAttention (2048 slices)