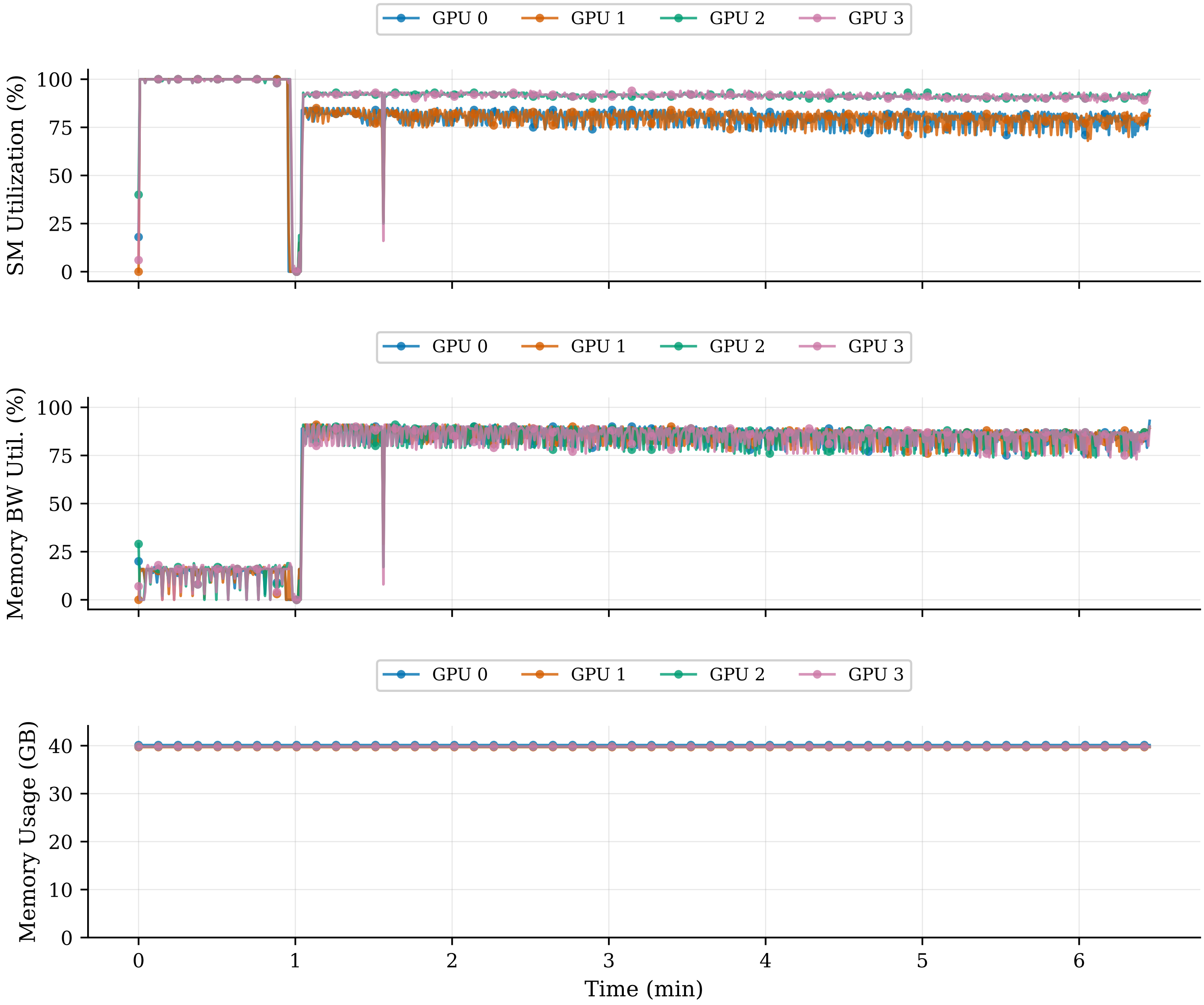
DeepSeek-R1-Distill-Llama-70B
TP=2, PP=4, Input=8192, Output=2048
Instance=2x g6.12xlarge, GPU=L4
Elapsed=387.4s, GPU Time=387.0s


GPU 0
GPU 1
GPU 2
GPU 3
SM Utilization (%)
GPU 0
GPU 1
GPU 2
GPU 3
Memory BW Util. (%)
GPU 0
GPU 1
GPU 2
GPU 3
Memory Usage (GB)
Time (min)