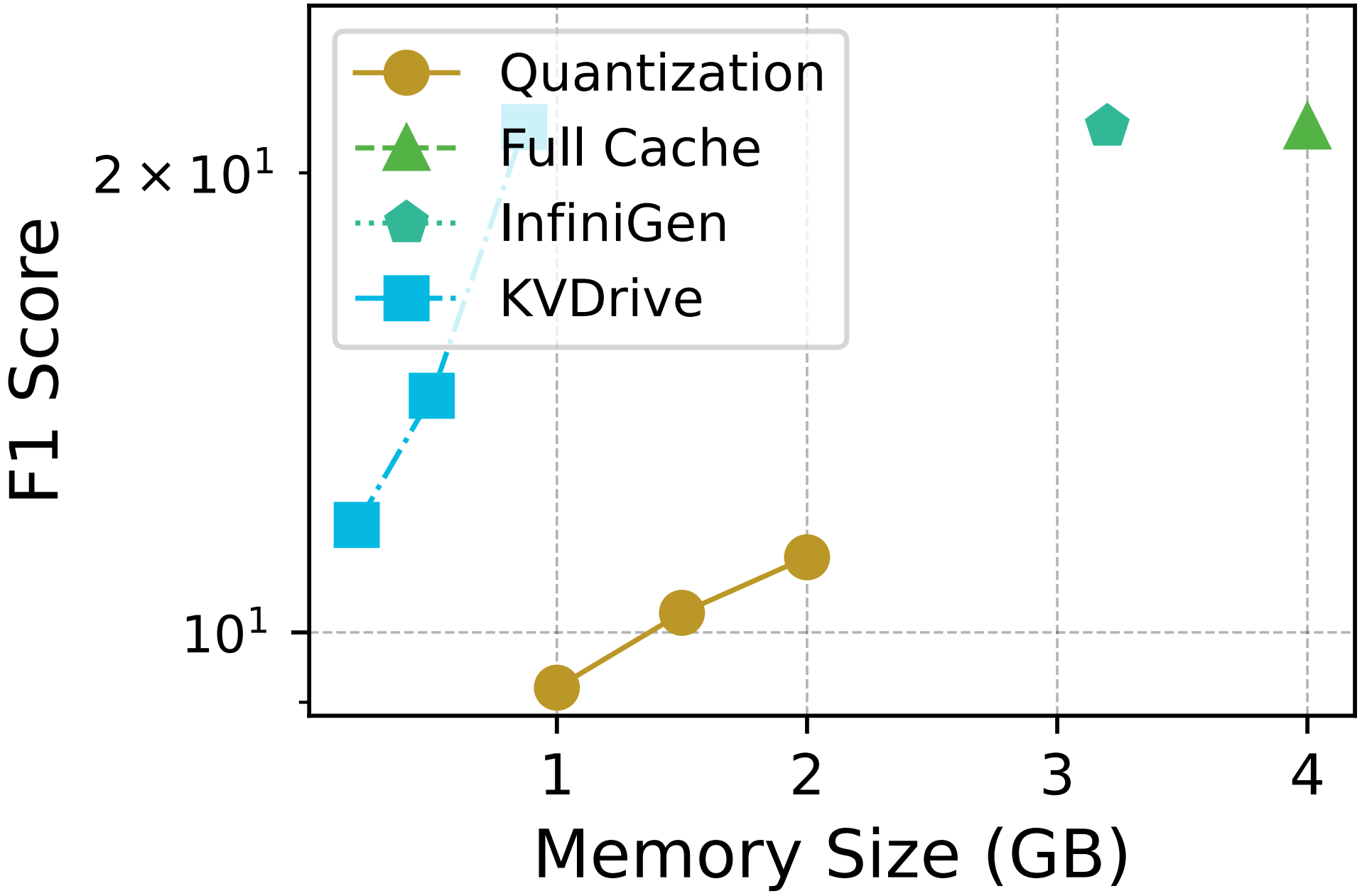
Quantization
Full Cache
InfiniGen
KVDrive
F1 Score
2×10^1
10^1
1
2
3
4
Memory Size (GB)