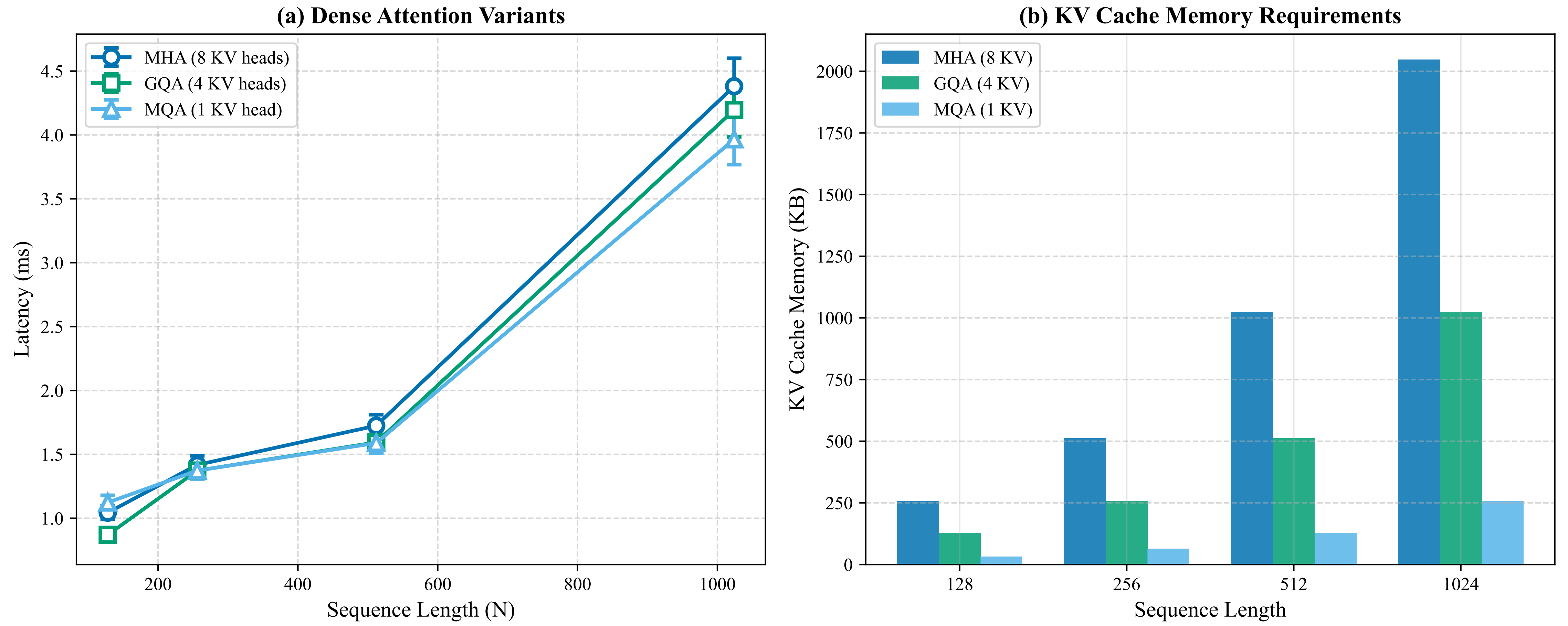
(a) Dense Attention Variants
MHA (8 KV heads)
GQA (4 KV heads)
MQA (1 KV head)
Latency (ms)
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
200
400
600
800
1000
Sequence Length (N)
(b) KV Cache Memory Requirements
MHA (8 KV)
GQA (4 KV)
MQA (1 KV)
KV Cache Memory (KB)
0
250
500
750
1000
1250
1500
1750
2000
128
256
512
1024
Sequence Length