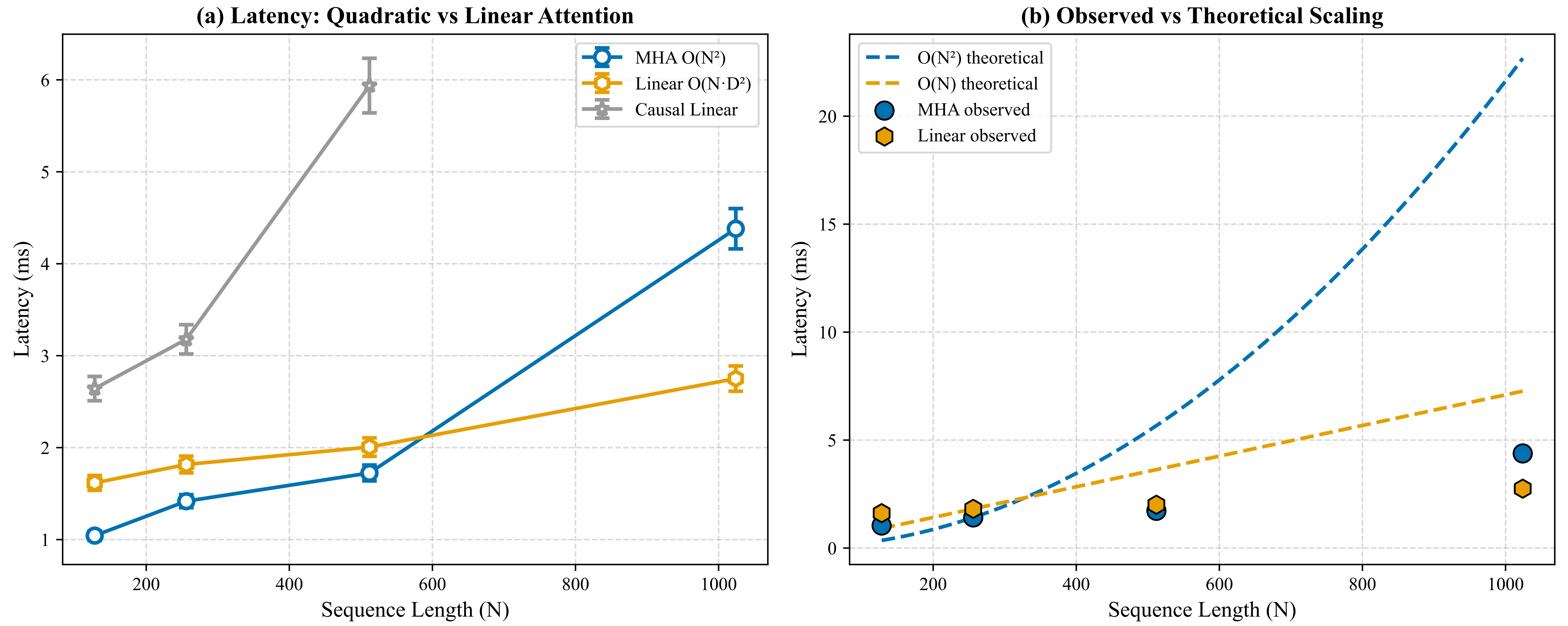
(a) Latency: Quadratic vs Linear Attention
MHA O(N²)
Linear O(N·D²)
Causal Linear
Latency (ms)
Sequence Length (N)
1
2
3
4
5
6
200
400
600
800
1000
(b) Observed vs Theoretical Scaling
O(N²) theoretical
O(N) theoretical
MHA observed
Linear observed
Latency (ms)
Sequence Length (N)
0
5
10
15
20
200
400
600
800
1000