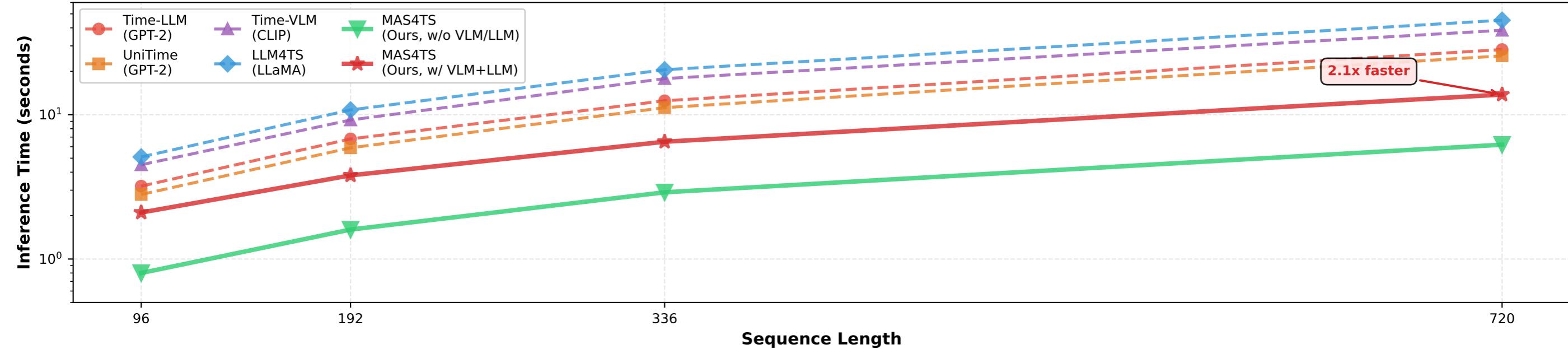
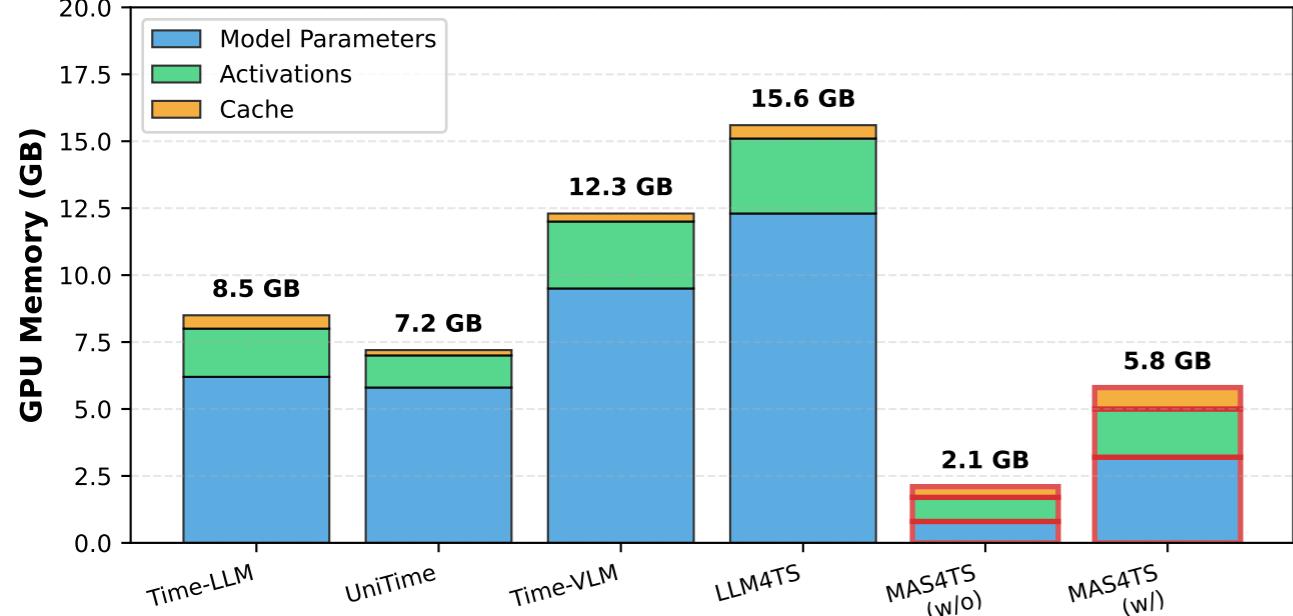


Efficiency Comparison: MAS4TS vs. Pre-trained LM-based Methods

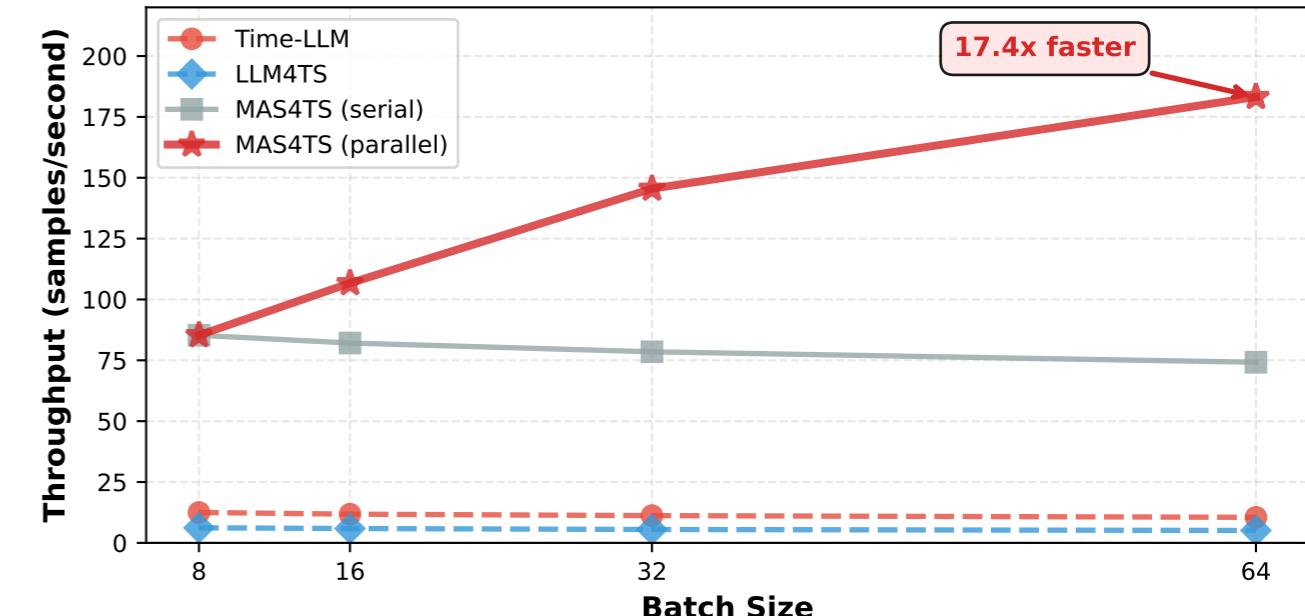
(a) Inference Time vs. Sequence Length



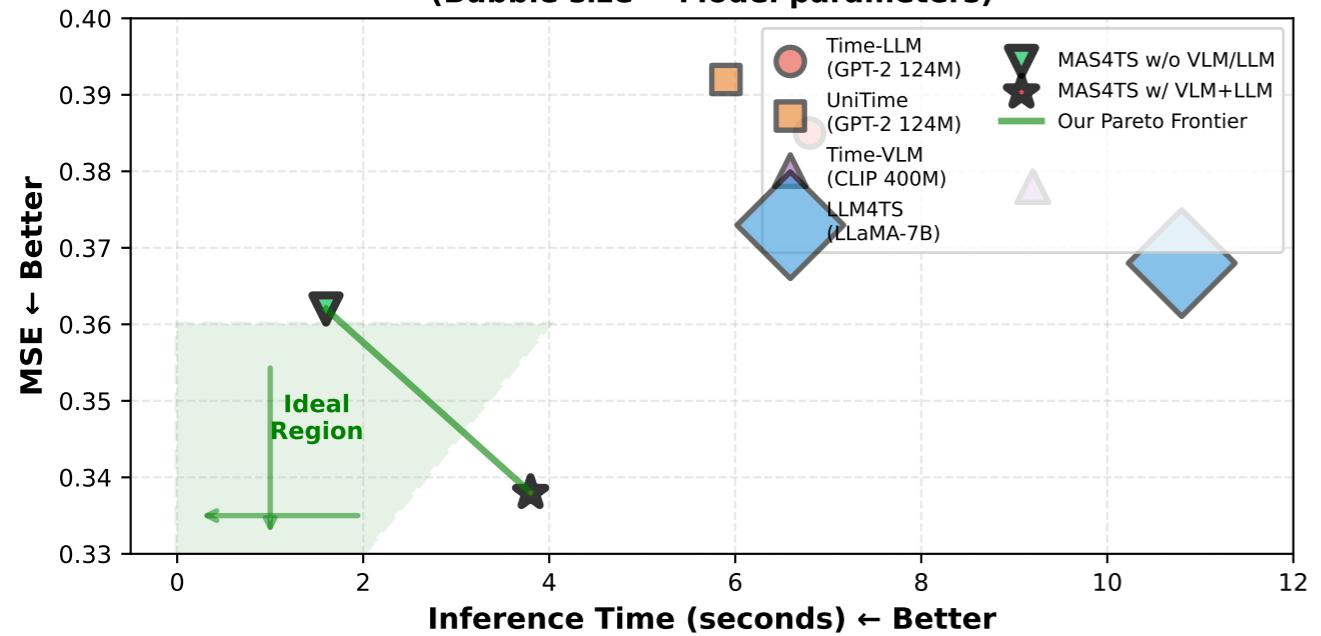
(b) Memory Consumption Breakdown



(c) Throughput Scaling with Batch Size



(d) Efficiency-Accuracy Trade-off
(Bubble size = Model parameters)



(e) Scalability: Time vs. Batch Size

