Figure 7: Real-time Performance (NVIDIA H100)

Legend:
- Latency (ms)
- Throughput (samples/sec)
- Real-time Threshold (10 ms)

Data points:
- Batch Size 1: 2000 samples/sec, 0.50ms latency
- Batch Size 8: 6667 samples/sec, 1.20ms latency
- Batch Size 32: 9143 samples/sec, 3.50ms latency
- Batch Size 64: 10667 samples/sec, 6.00ms latency