Inference Latency Comparison Across Scenarios