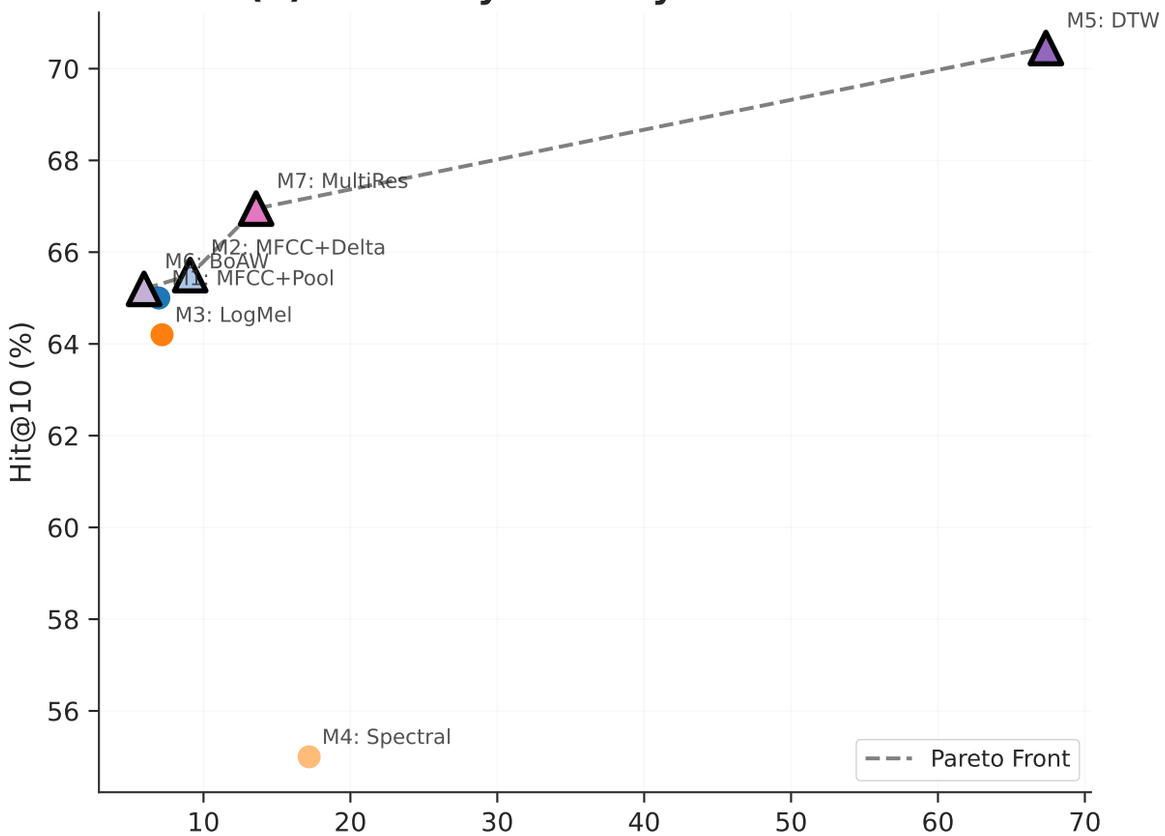
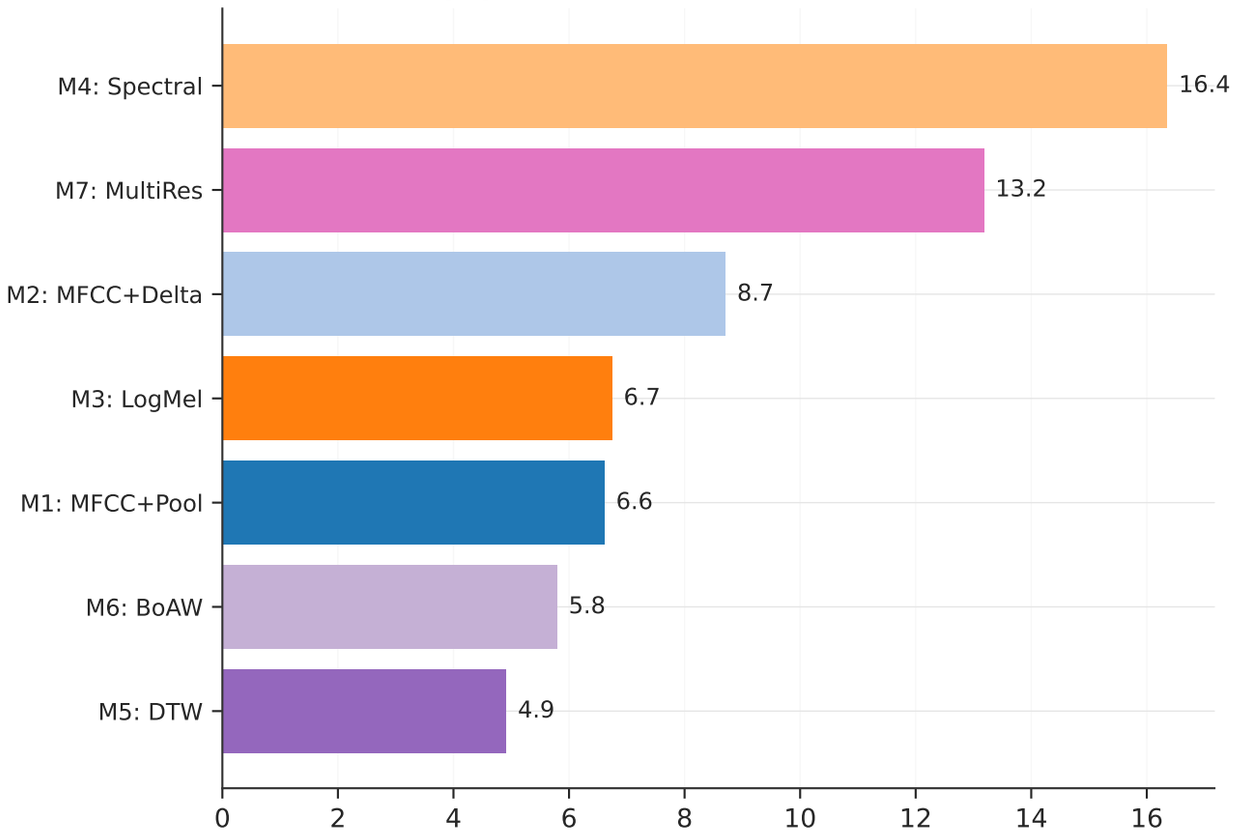
Efficiency Analysis

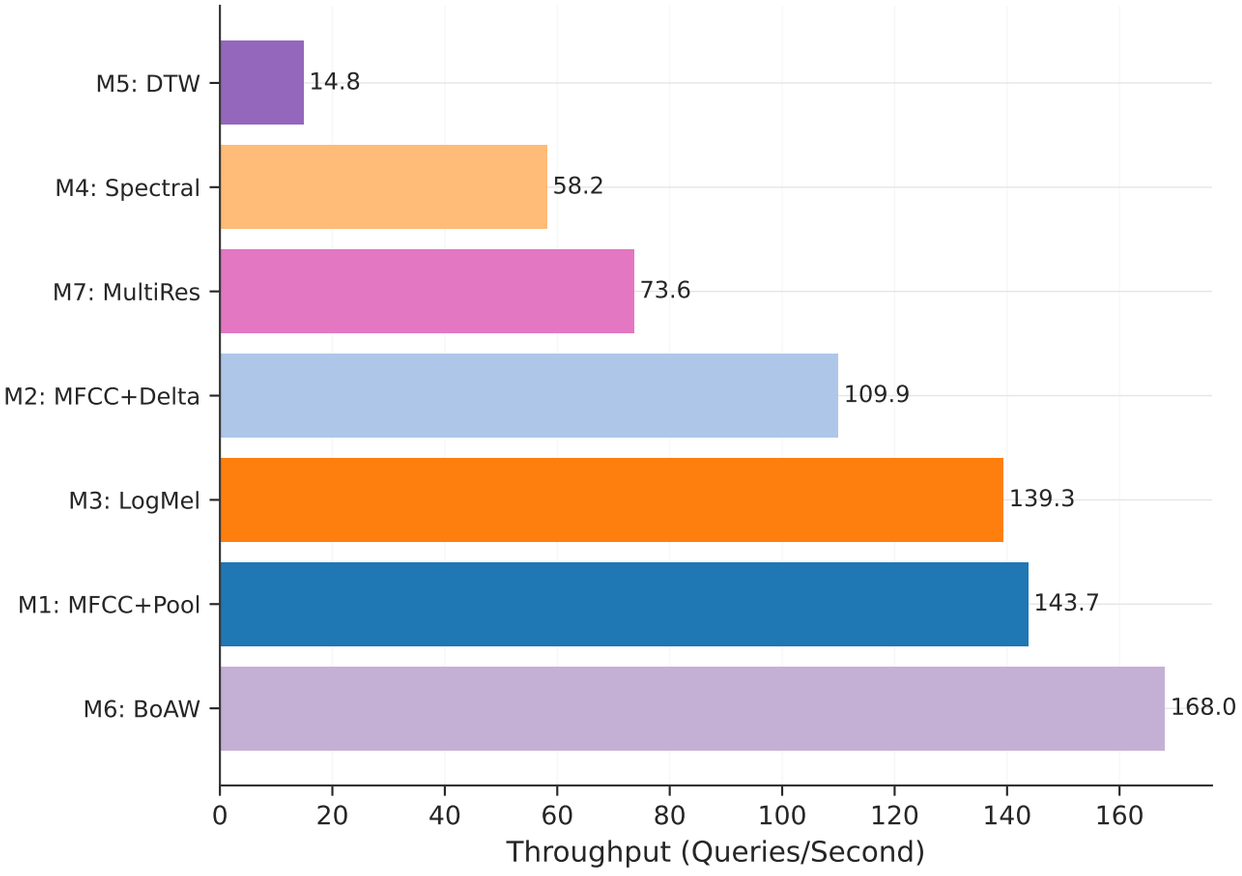
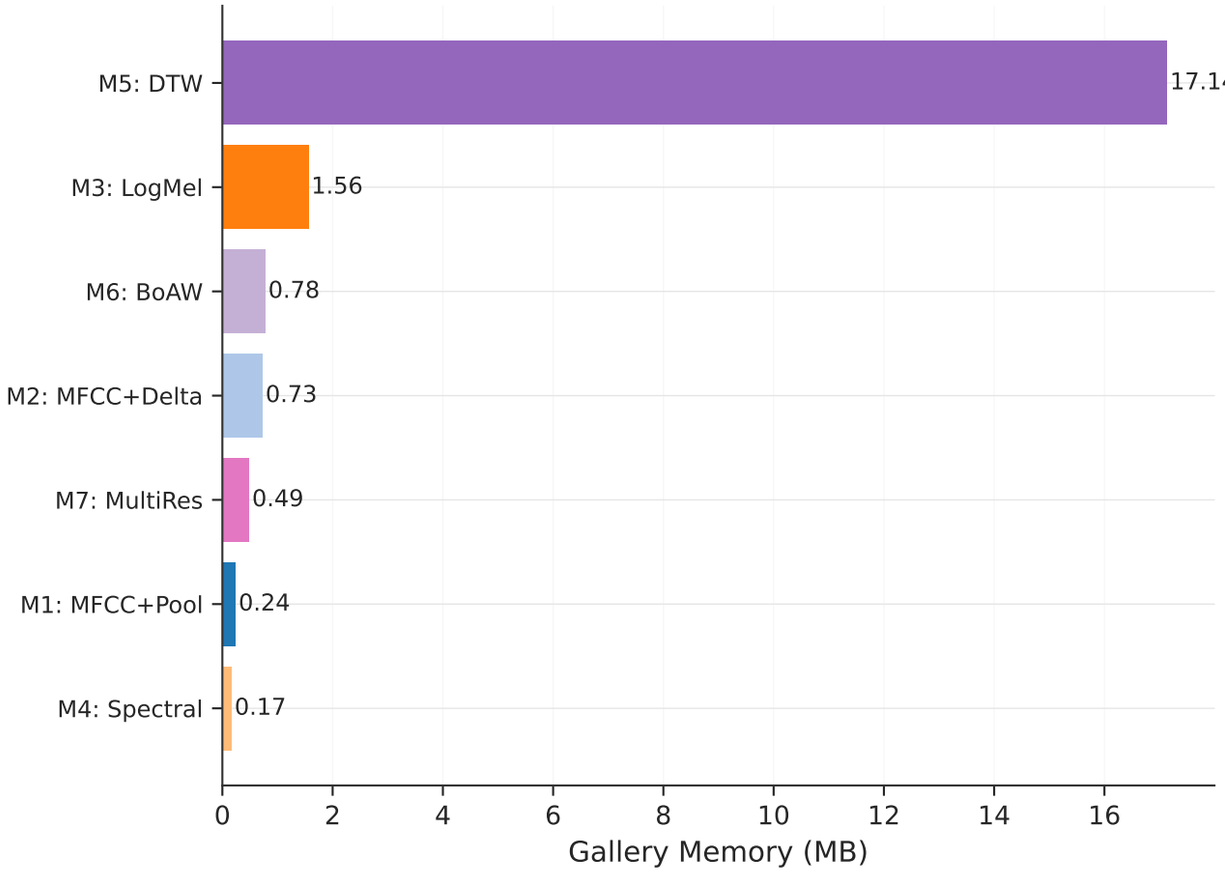
(a) Accuracy-Latency Pareto Front



(b) Feature Extraction Time



(c) Query Throughput



(d) Memory Footprint