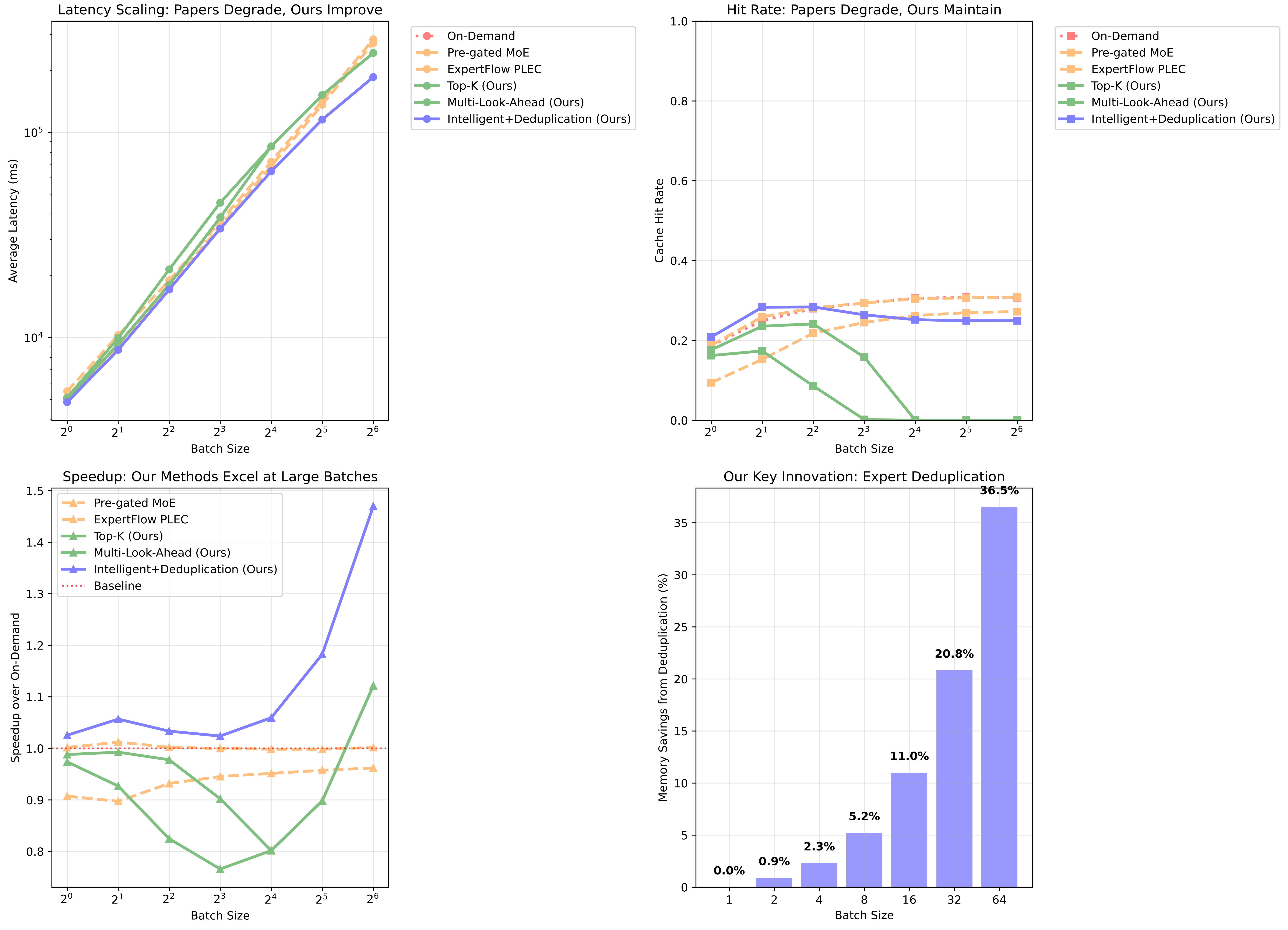
Paper Baselines vs Our Batch-Aware Contributions


Latency Scaling: Papers Degrade, Ours Improve
Average Latency (ms)
Batch Size
On-Demand
Pre-gated MoE
ExpertFlow PLEC
Top-K (Ours)
Multi-Look-Ahead (Ours)
Intelligent+Deduplication (Ours)
Hit Rate: Papers Degrade, Ours Maintain
Cache Hit Rate
Speedup: Our Methods Excel at Large Batches
Speedup over On-Demand
Baseline
Our Key Innovation: Expert Deduplication
Memory Savings from Deduplication (%)