## **Paper Baselines vs Our Batch-Aware Contributions** Latency Scaling: Papers Degrade, Ours Improve Hit Rate: Papers Degrade, Ours Maintain On-Demand On-Demand Pre-gated MoE --- Pre-gated MoE ExpertFlow PLEC ExpertFlow PLEC Top-K (Ours) Top-K (Ours) Multi-Look-Ahead (Ours) 8.0 Multi-Look-Ahead (Ours) Intelligent+Deduplication (Ours) Intelligent+Deduplication (Ours) $10^{5}$ Average Latency (ms) Cache Hit Rate 9.0 9.0 $10^{4}$ 0.0 2<sup>'3</sup> 2<sup>5</sup> 2<sup>2</sup> 2<sup>4</sup> 2<sup>2</sup> 21 Batch Size Batch Size Speedup: Our Methods Excel at Large Batches Our Key Innovation: Expert Deduplication → Pre-gated MoE ExpertFlow PLEC 35 Top-K (Ours) Multi-Look-Ahead (Ours) Intelligent+Deduplication (Ours) Memory Savings from Deduplication (%) 0 1 2 0 0 2 1 0 0 0 10 Baseline 1.3 Speedup over On-Demand 1.1 0.0 20.8% 11.0% 0.9 5.2% 5 2.3% 8.0 0.9% 0.0% 2<sup>'3</sup> 26 2<sup>2</sup> $2^1$ 2<sup>4</sup> 2<sup>5</sup> 8 16 32 64 Batch Size Batch Size