



ArkVale: Efficient Generative LLM Inference with Recallable Key-Value Eviction

Renze Chen¹, Zhuofeng Wang¹, Beiquan Cao¹, Tong Wu¹, Size Zheng¹, Xiuhong Li¹, Xuechao Wei¹, Shengen Yan², Meng Li¹, Yun Liang¹

¹Peking University ²Infinigence-AI

E-mail: crz@pku.edu.cn

Code: <https://github.com/pku-liang/ArkVale>

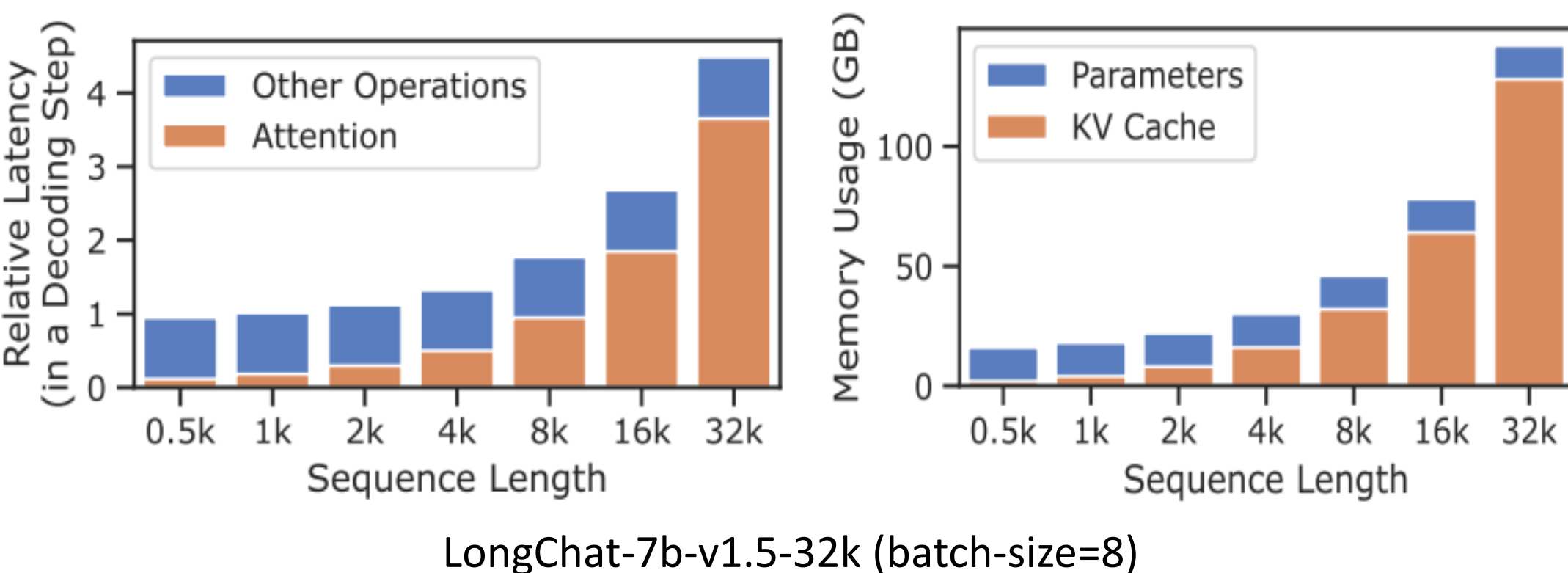


Scan to access our code

1. Motivation

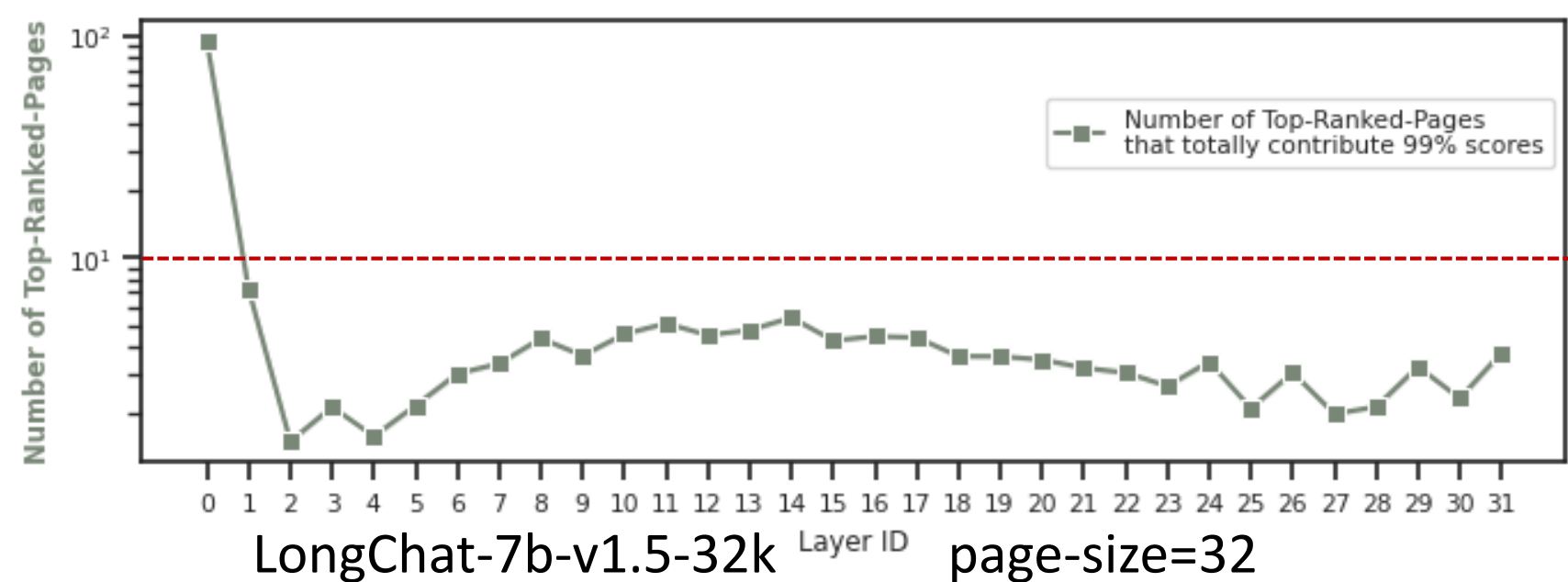
Impact of Long-context

- Long context attention can be the bottleneck of LLM decoding
- Long context can be the memory bottleneck of LLM decoding, which hampers the use of larger batch-size for serving.

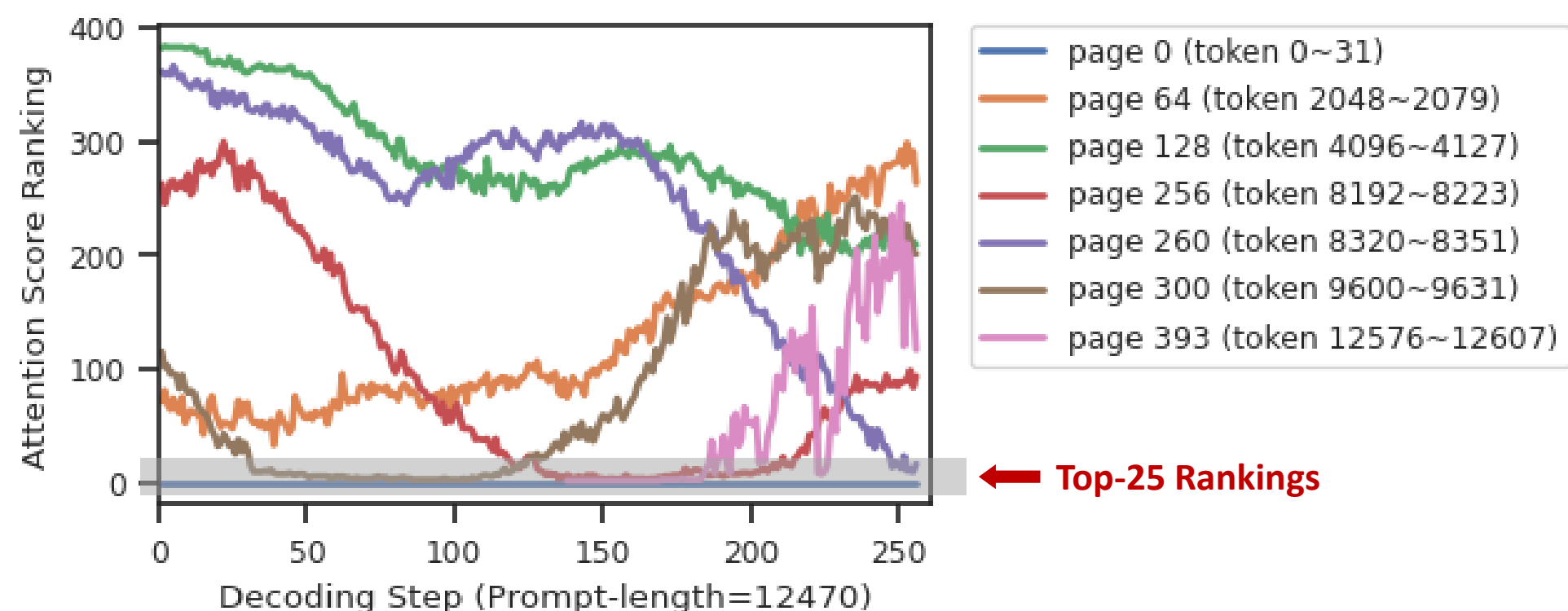


2. Observation

In most layers, less than 10 KV-cache pages contributing most attention scores.

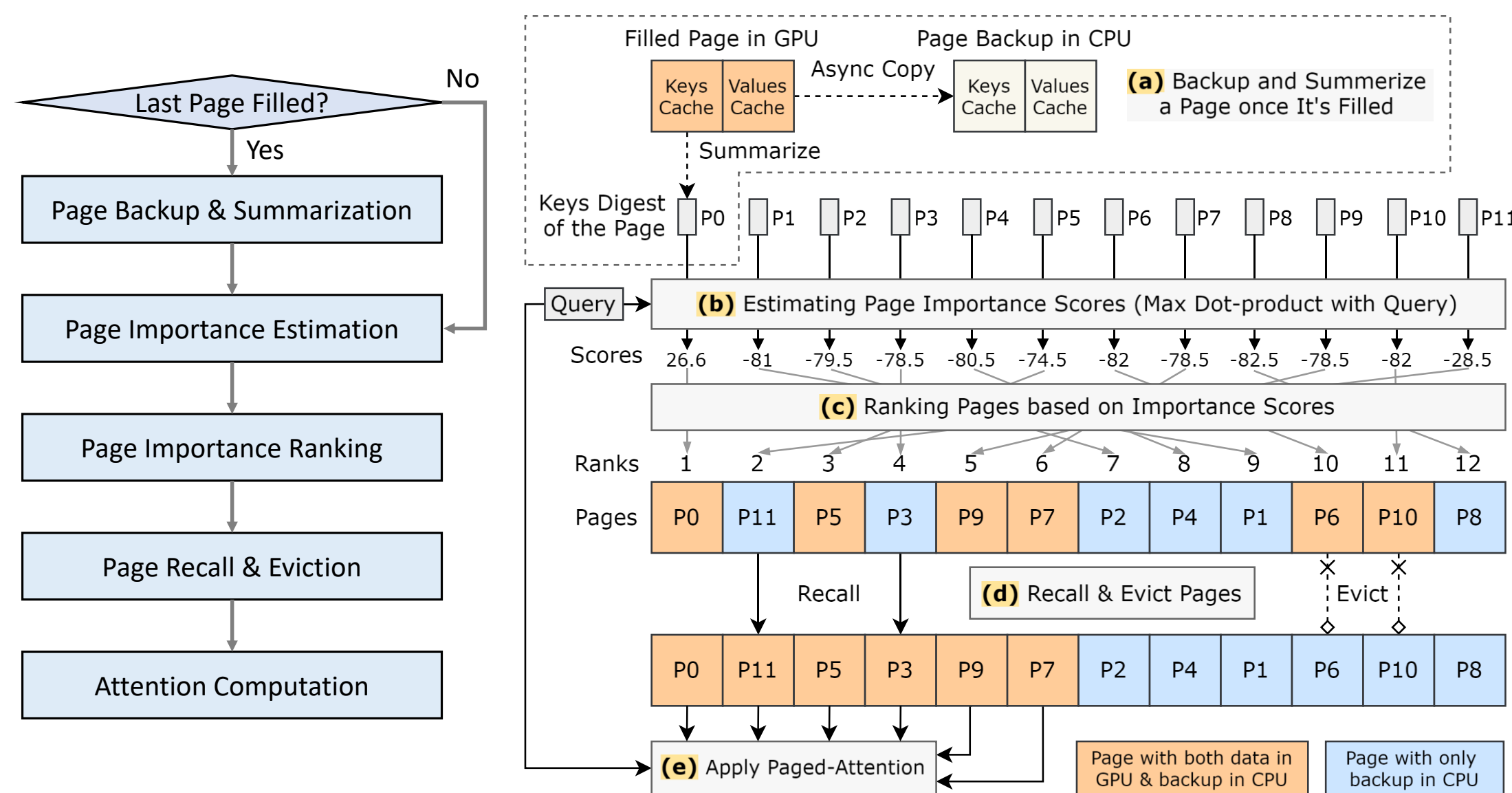


Importance of KV-cache token/page can dynamically change overtime



3. Techniques

Workflow Overview



Page Summarization & Importance Estimation using Bounding-volume

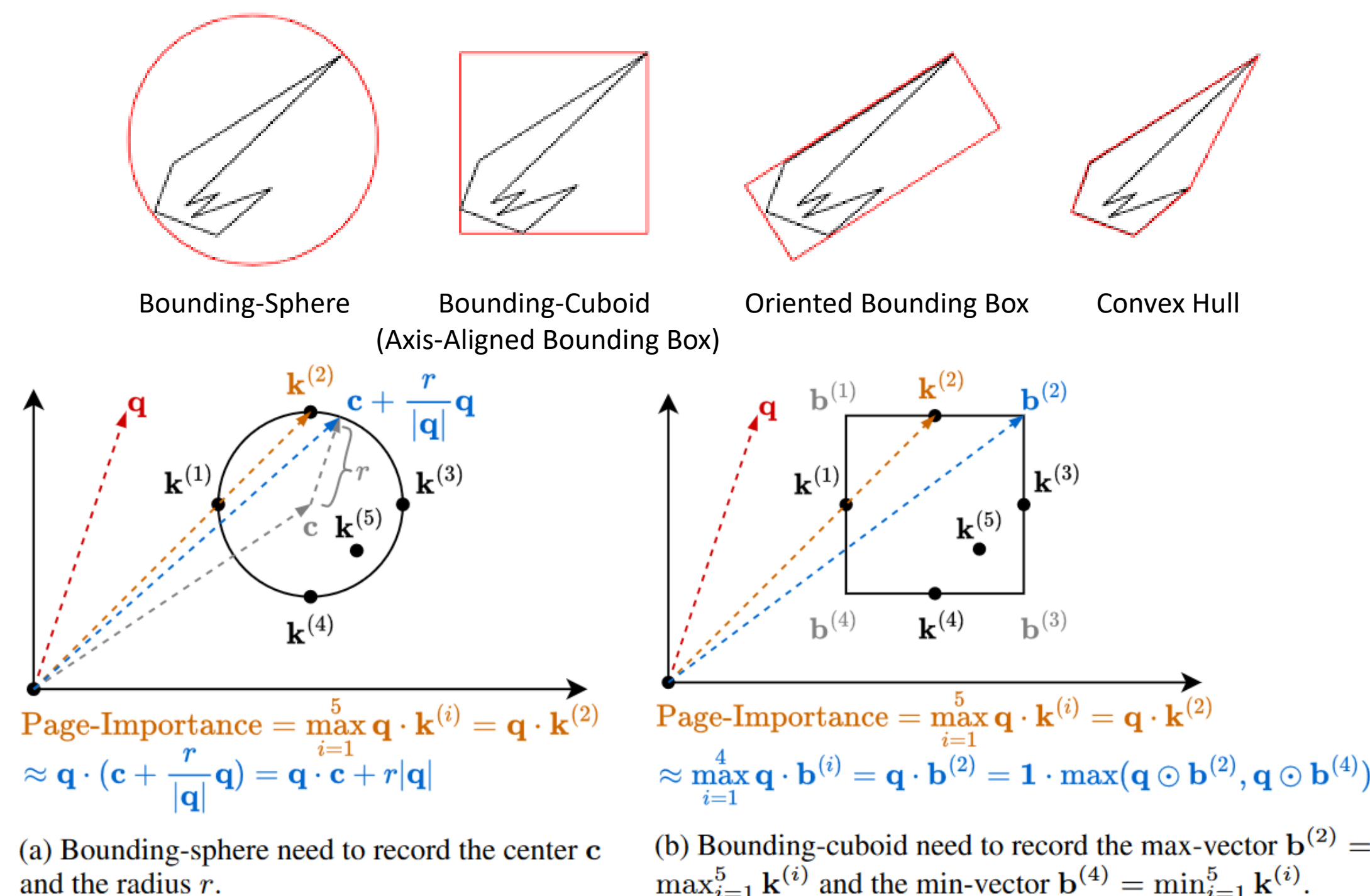
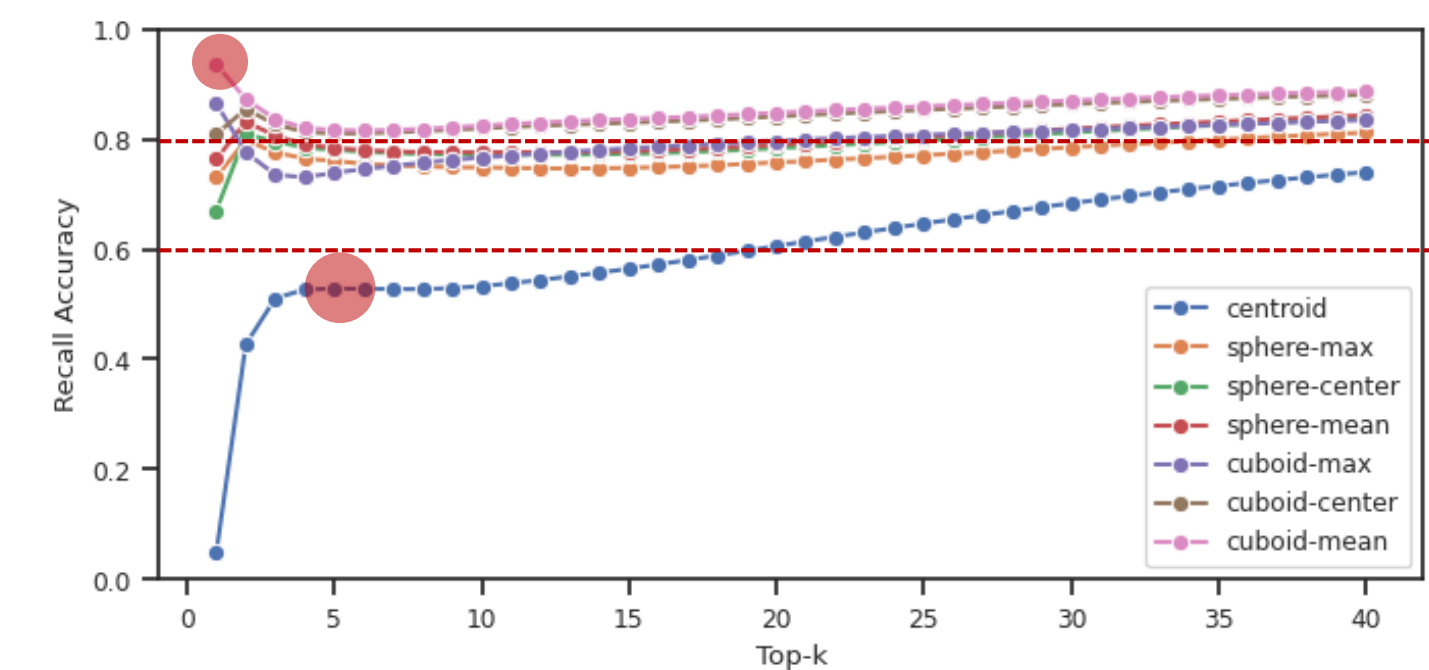


Figure 6: Summarize page keys $\{\mathbf{k}^{(i)}\}_{i=1}^5$ into their bounding-volume (sphere/cuboid). We can estimate the max-dot-product between query \mathbf{q} and keys $\{\mathbf{k}^{(i)}\}_{i=1}^5$ using the bounding-volume.

4. Evaluation

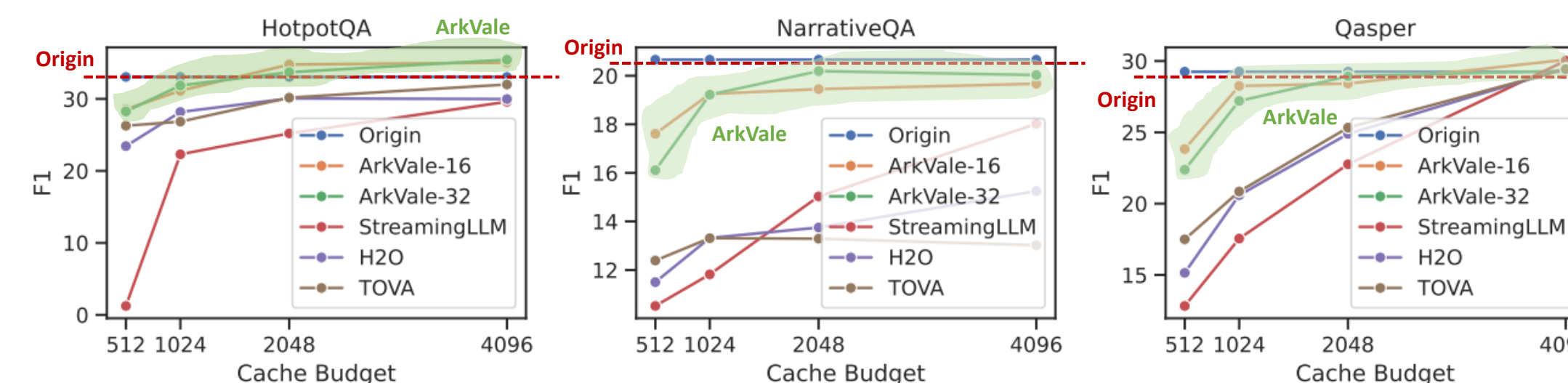
Top-k recall accuracy of different importance estimation methods

- Baseline method (centroid) cannot achieve even 60% top-5 recall accuracy.
- Our cuboid-mean method ensure 95% top-1 recall accuracy, and can achieve 80% top-k recall accuracy for all k .



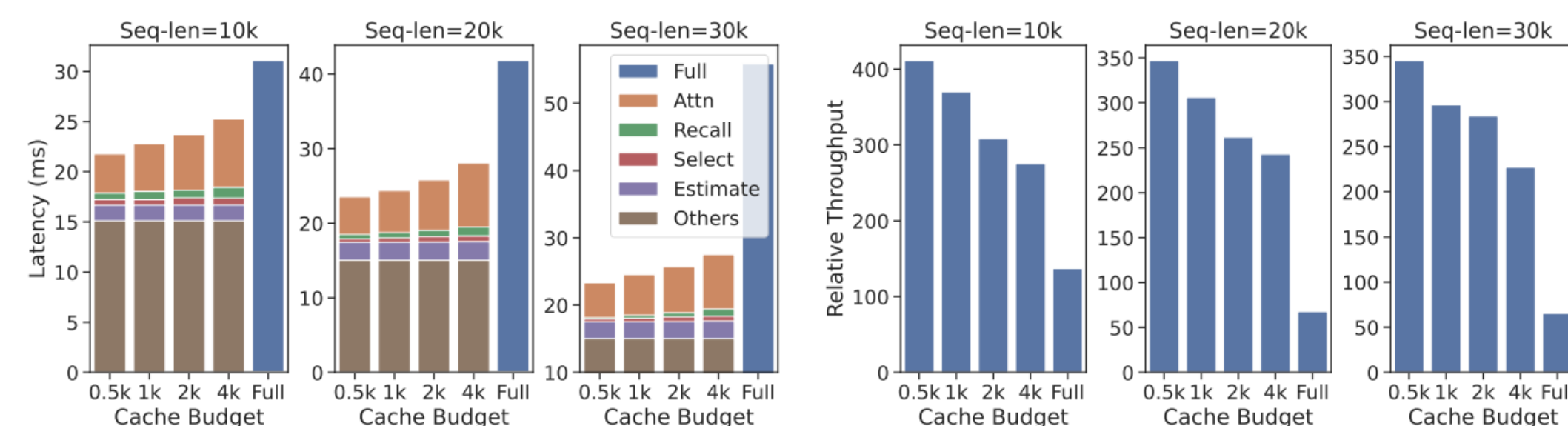
Part of Evaluation Results on Long-Bench

- ArkVale can surpass all baselines with different datasets and cache-budgets.
- ArkVale can approach or even surpass "Origin".
- ArkVale-16 (page-size=16) usually outperforms ArkVale-32 (page-size=32).



Decoding Latency & Throughput Evaluation

- Allocate 40 GB GPU memory for KV-cache (and page digests) on A100 GPU.
- Compared to baseline, ArkVale can achieve up to 2.2x decoding speedup.
- Compared to baseline, ArkVale can achieve up to 6x decoding throughput.



(a) Latency Breakdown (batch-size=4)

(b) Throughput Comparison