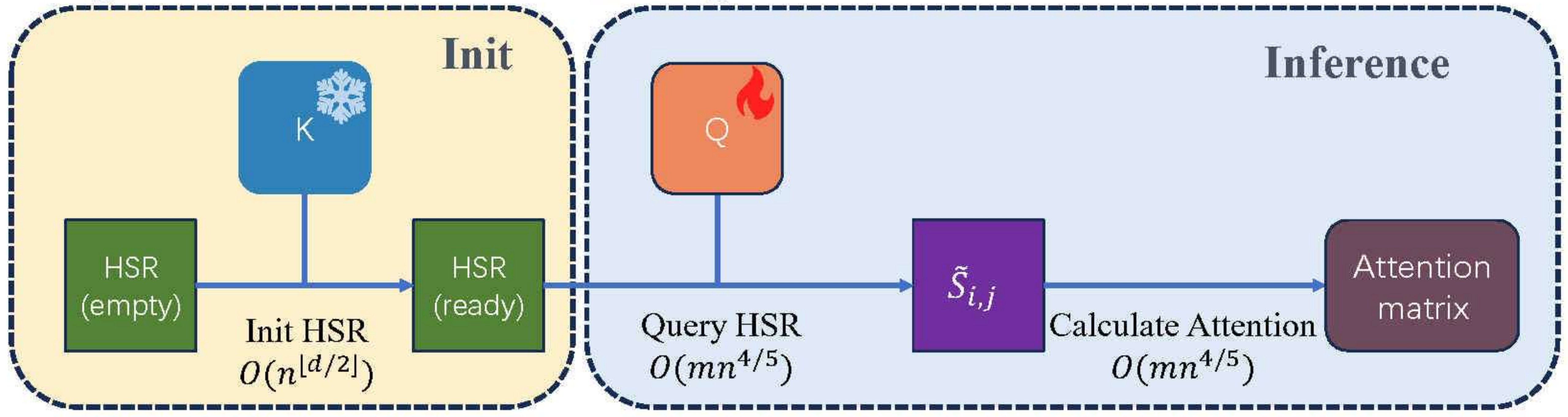
HSR-enhanced sparse attention acceleration

Bo Chen, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song

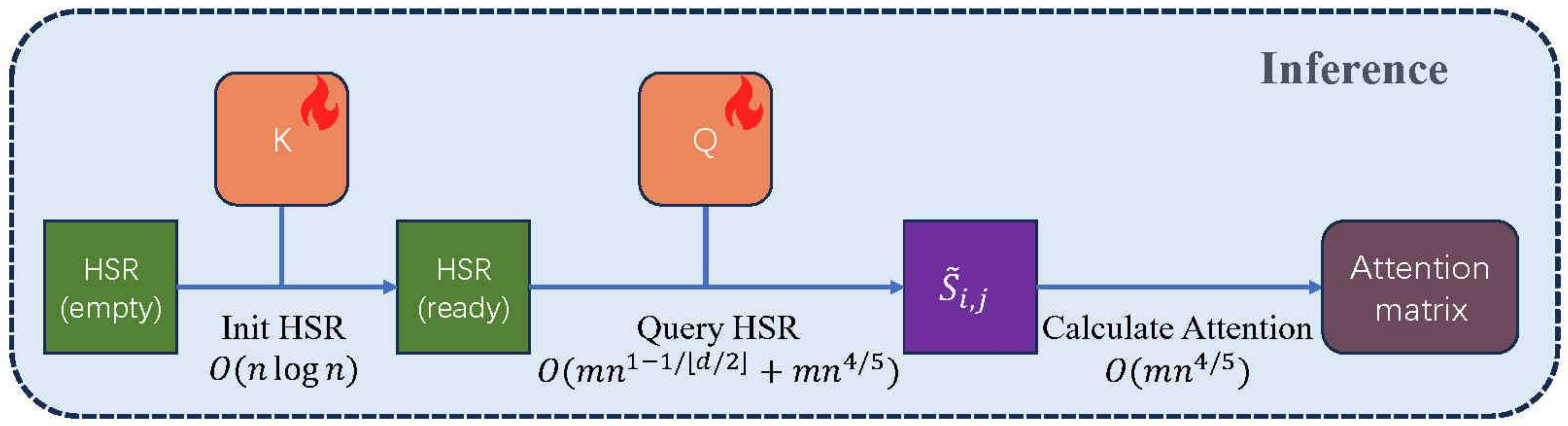
ReLU attention running too slow?

Using HSR data-structure to accelerate it!

Algorithm 1: Generation Decoding



Algorithm 2: Prompt Prefilling



- Half Space Range Reporting data structure [AEM92]

Part 1 $\mathcal{T}_{\mathsf{init}}(n,d)=O_d(n\log n)$ $\mathcal{T}_{\mathsf{query}}(n,d,k)=O(dn^{1-1/\lfloor d/2\rfloor}+dk)$

Part 2 $\mathcal{T}_{\mathsf{init}}(n,d)=O(n^{\lfloor d/2\rfloor})$ $\mathcal{T}_{\mathsf{query}}(n,d,k)=O(d\log(n)+dk)$

Theoretical Results

Theorem 1 (Accelerating ReLU attention generation decoding)
Our algorithm can accelerate the generate decoding of ReLU attention based Tranformers from $O(mn)$ to $O(mn^{4/5})$ with high probability , where m denotes the generated length and n denotes the prefilled length. ($n \gg m$).

Theorem 2 (Accelerating ReLU attention prompt prefilling)
Our algorithm can accelerate the prompt prefilling of ReLU attention based Tranformers from $O(n^2)$ to $O(n^{2-1/\lfloor d/2\rfloor}+n^{1+4/5})$ with high probability , where n denotes the prefilled prompt length.

Extensions Our HSR based method can easily extend from ReLU attention to standard Softmax attention. In Softmax case, we only calculate the attention matrix values within the “massive activated index set”. Those “massive activated” values can greatly approximate the actual value while saving great computation time.

Take-Home Message The sparsity within the attention mechanism can be used to accelerate the attention mechanism itself. In this work, we first analyze the ReLU attention, showing that the HSR data structure can accelerate it. Then, we extend our results to conventional Softmax attention setting, which shows the generalization of our method.