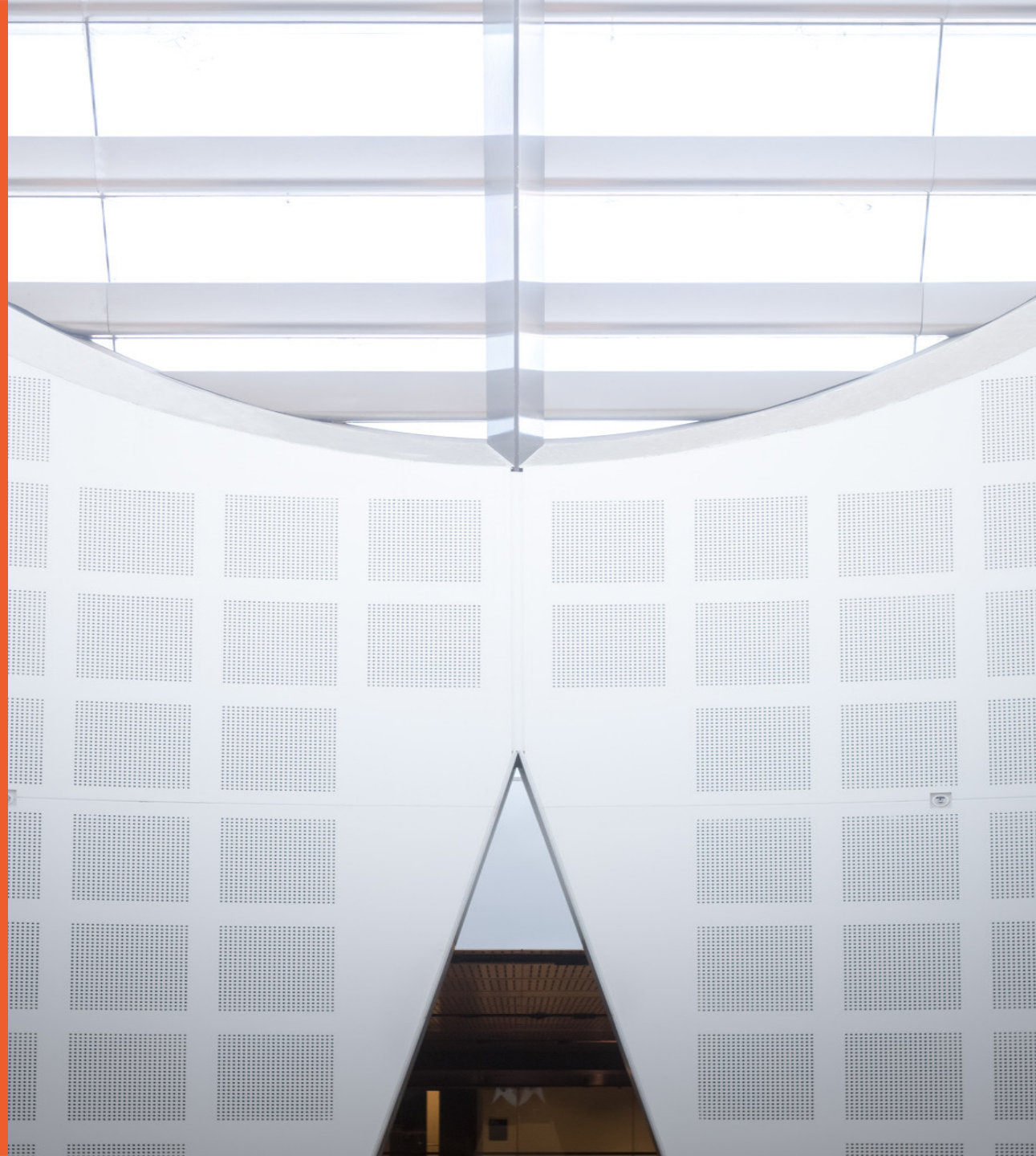
Seminar 2: Efficiency of LLM

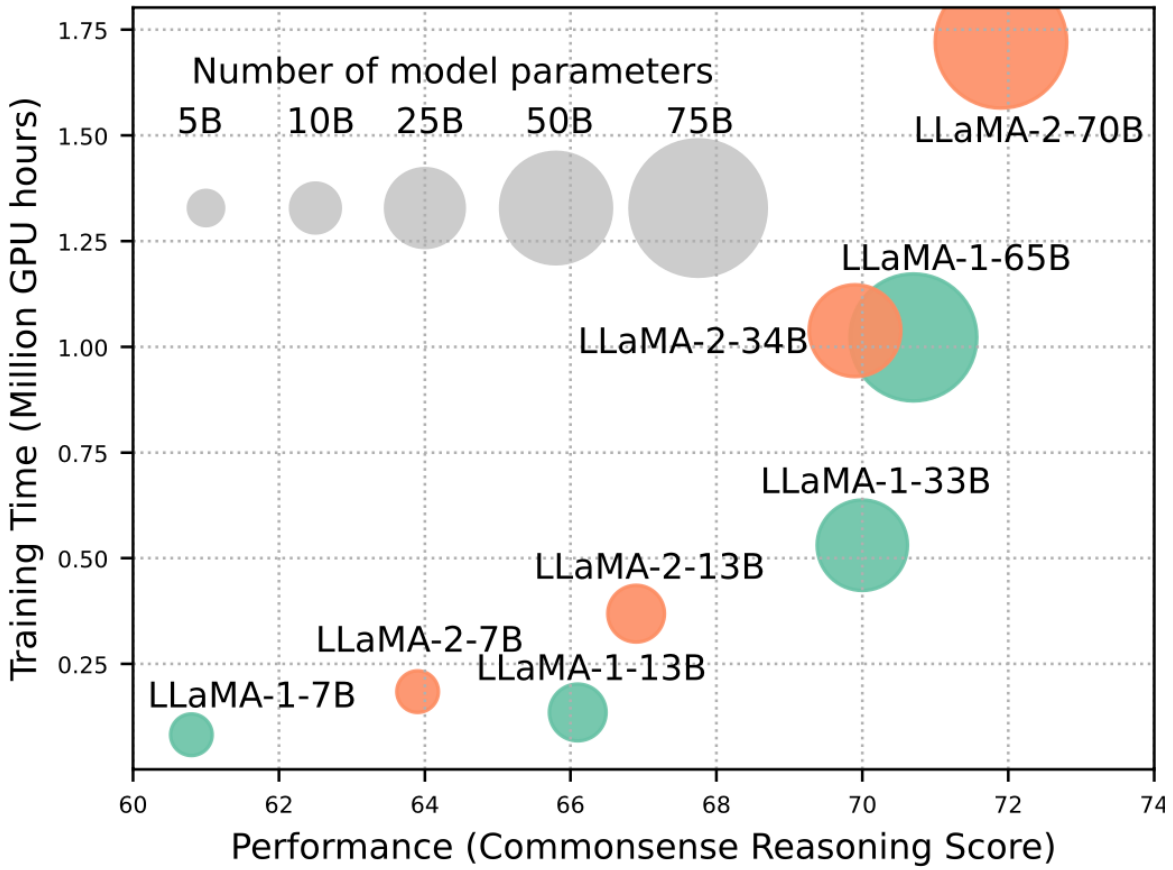
Xiaohuan Pei

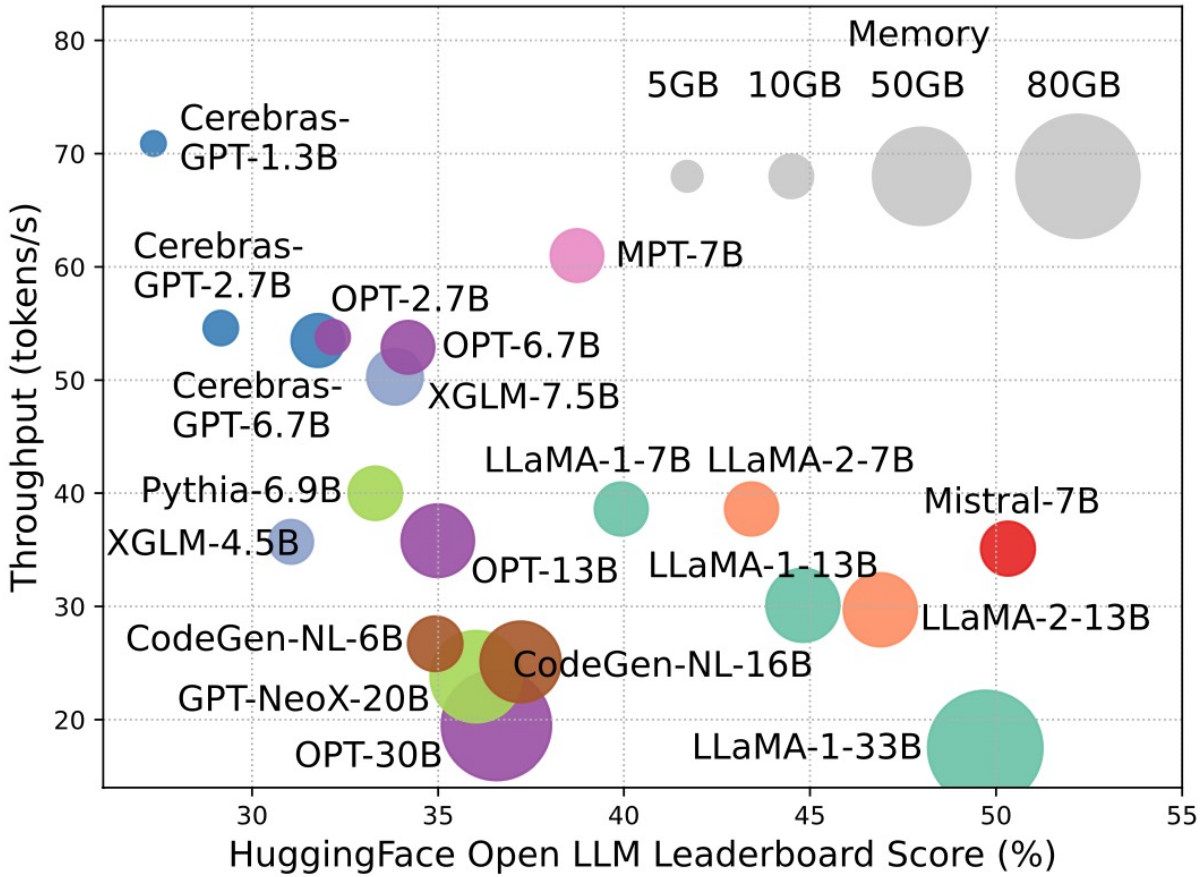
xiaohuan.pei@sydney.edu.au
School of Computer Science

THE UNIVERSITY OF SYDNEY

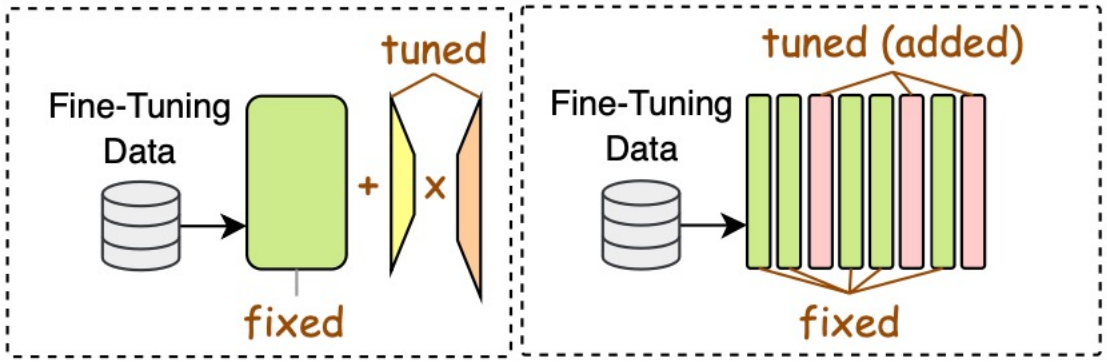


How important of efficiency ?

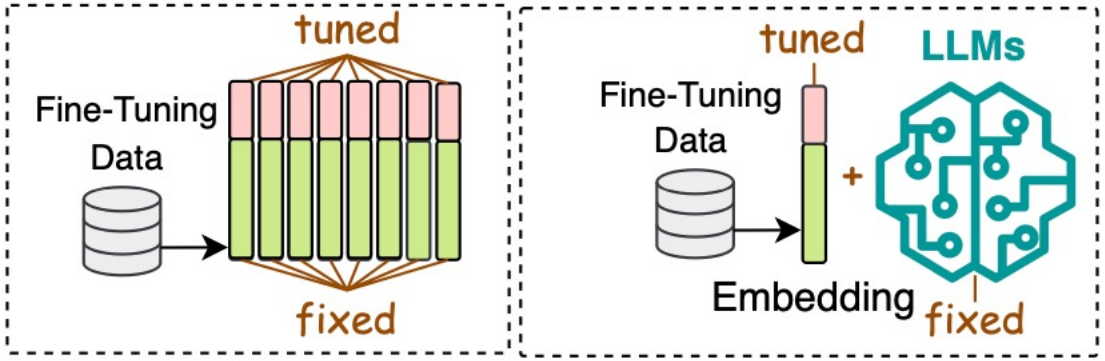




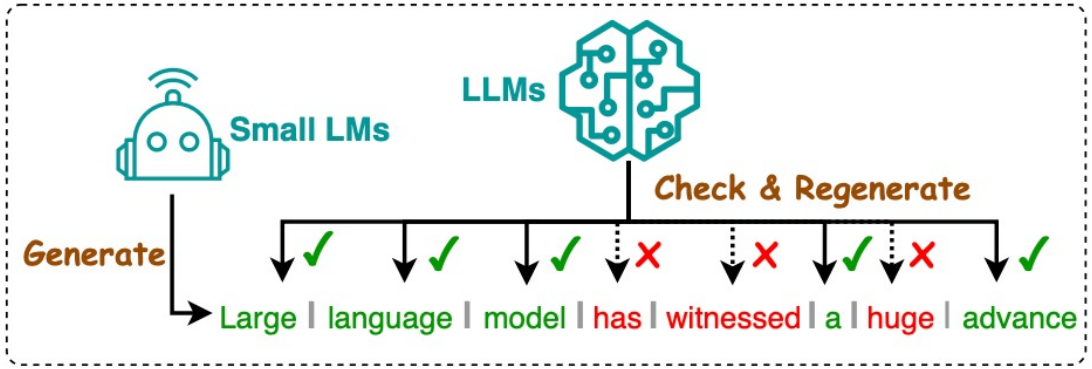
Overview of Efficiency in LLM



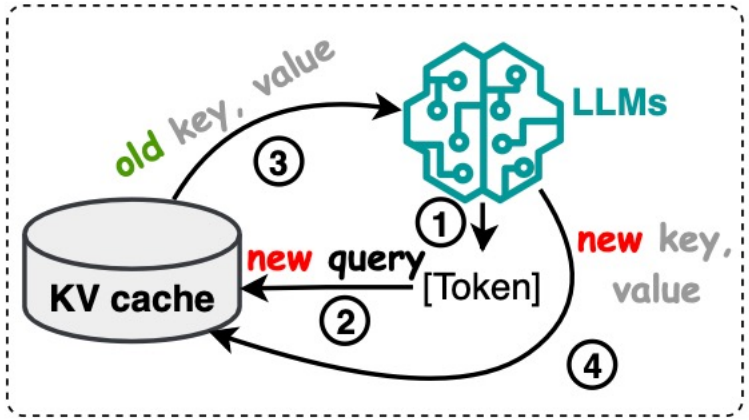
(a) Low-Rank Adaptation (b) Adapter-based Tuning



(c) Prefix Tuning (d) Prompt Tuning

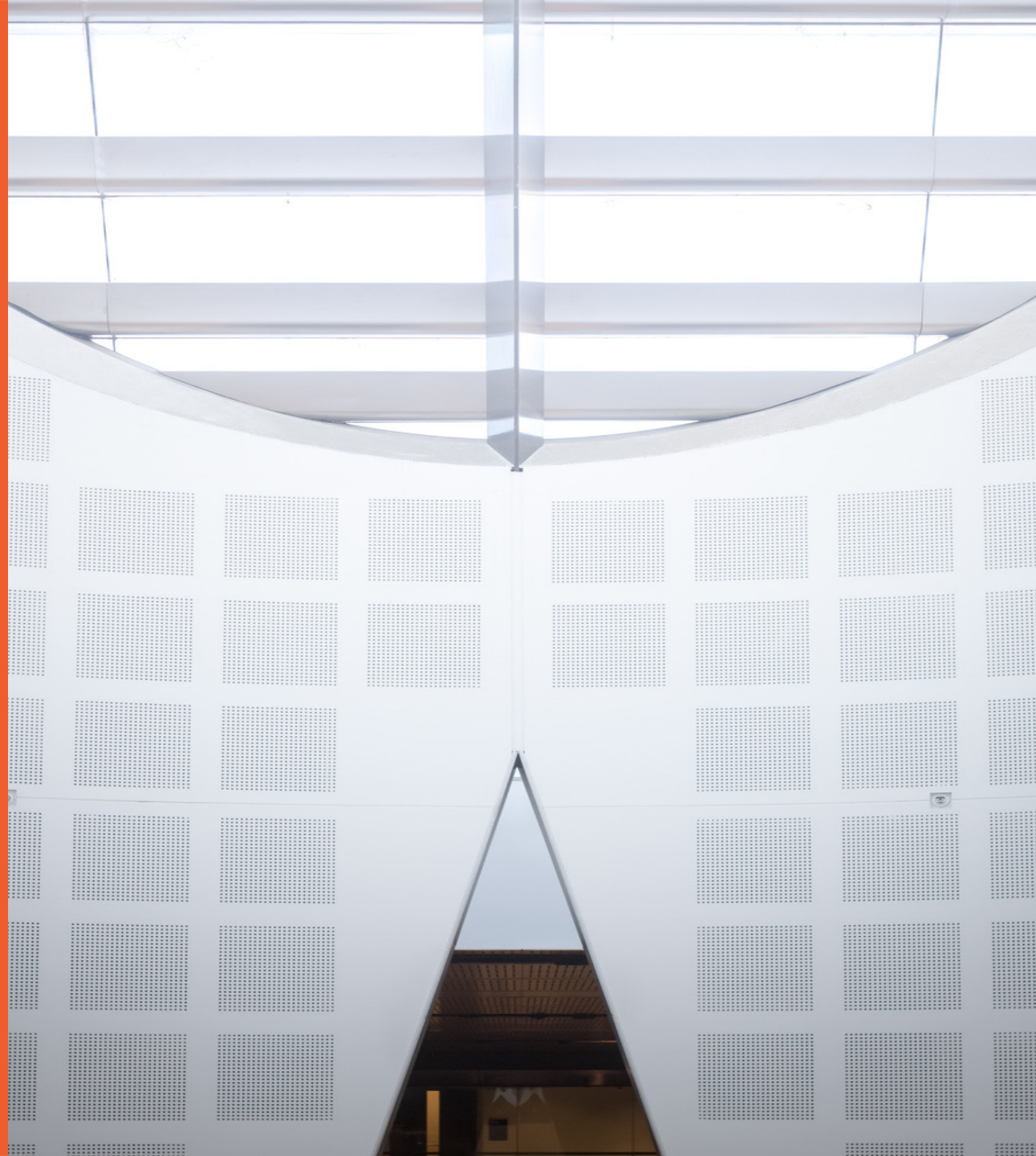


(a) Speculative Decoding



(b) KV-Cache Optimization

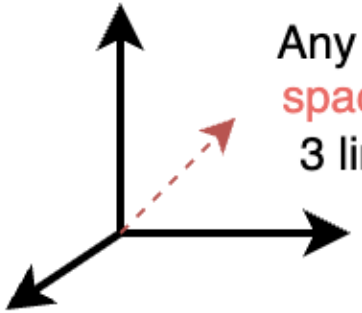
Seminar 2, Part 1.1: Efficiency of Fune-Tuning



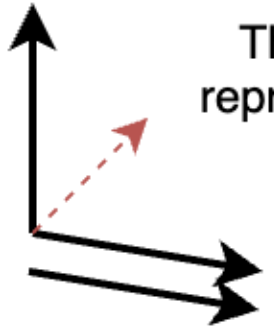
THE UNIVERSITY OF SYDNEY

Low-Rank Adaption: LoRA

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}, \quad \text{Rank}(\mathbf{A}) = 3. \qquad \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 2 & 4 & 6 \end{bmatrix}, \quad \text{Rank}(\mathbf{A}) = 2.$$



Any vectors in 3-dimensional space could be represent by 3 linear dependent vectors

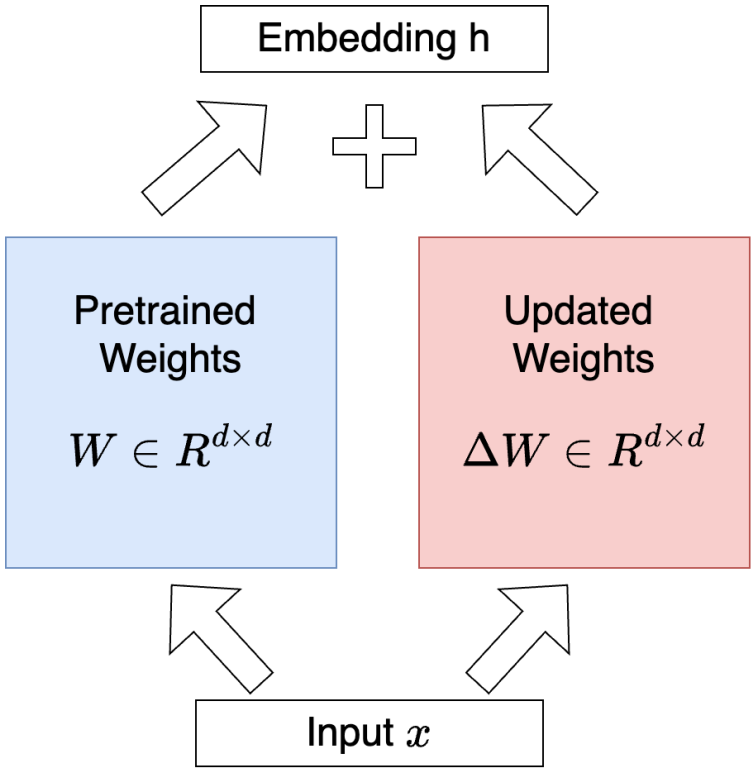


The vector could **not** be represented by only 2 linear dependent vectors

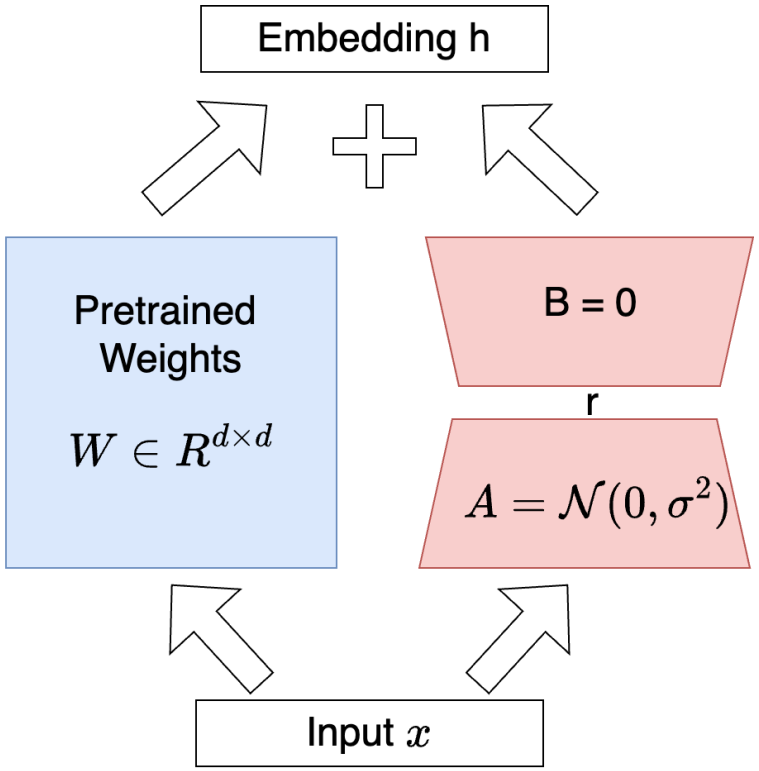
We hope the rank is large in weights as it can represent more dimentional latent features

The weights pretrained only need specfic base vectors for specific tasks

Low-Rank Adaption: LoRA

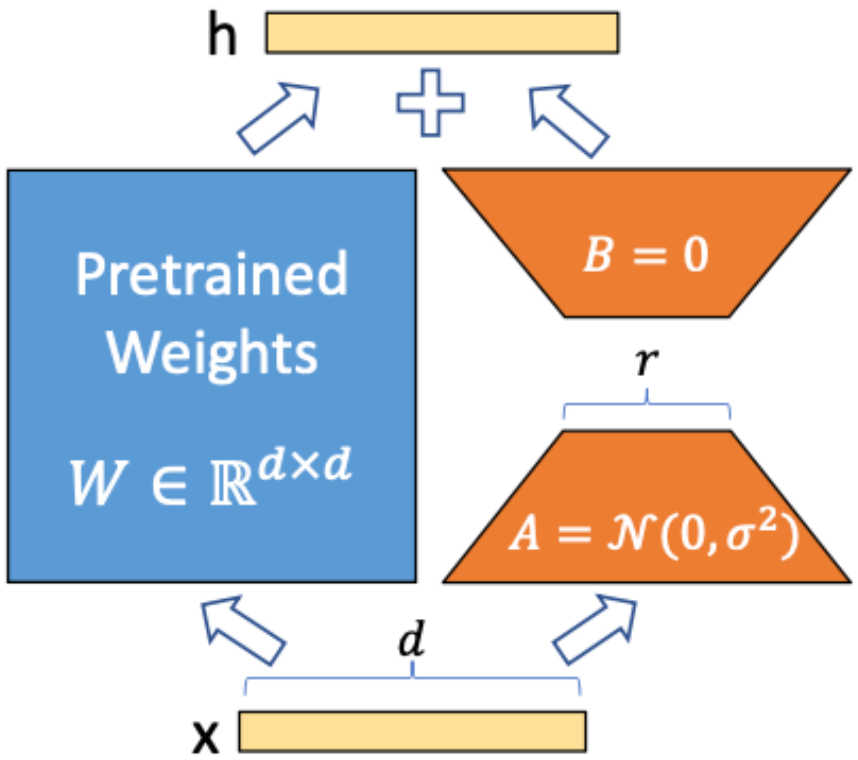


$$h = W_0 x + \Delta W x$$



$$h = W_0 x + BAx$$

Low-Rank Adaption: LoRA



$$h = W_0 x + \Delta W x = W_0 x + BAx$$

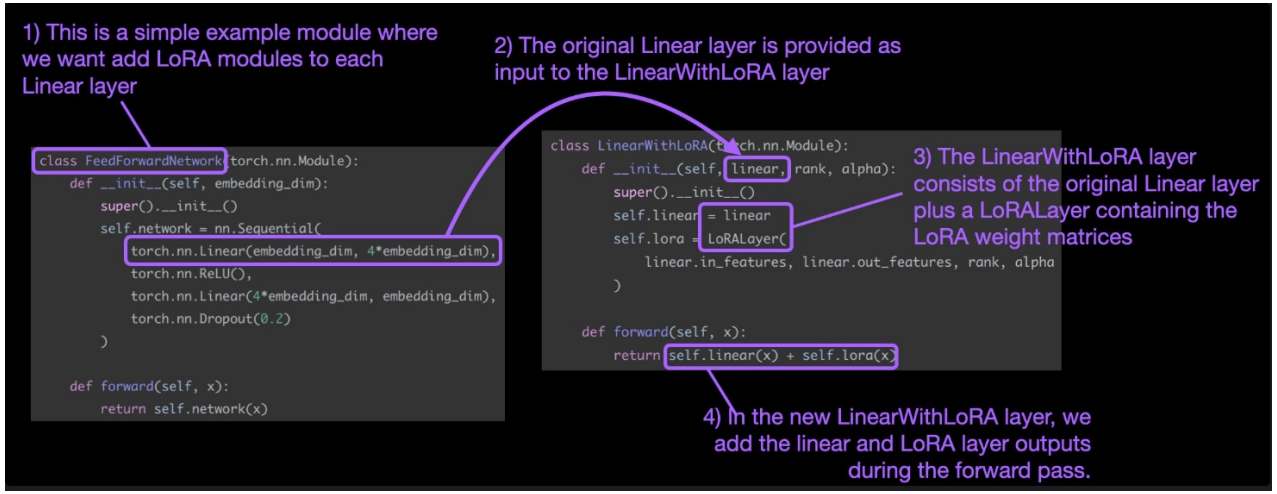
Why B = 0

Tuning starts from original W

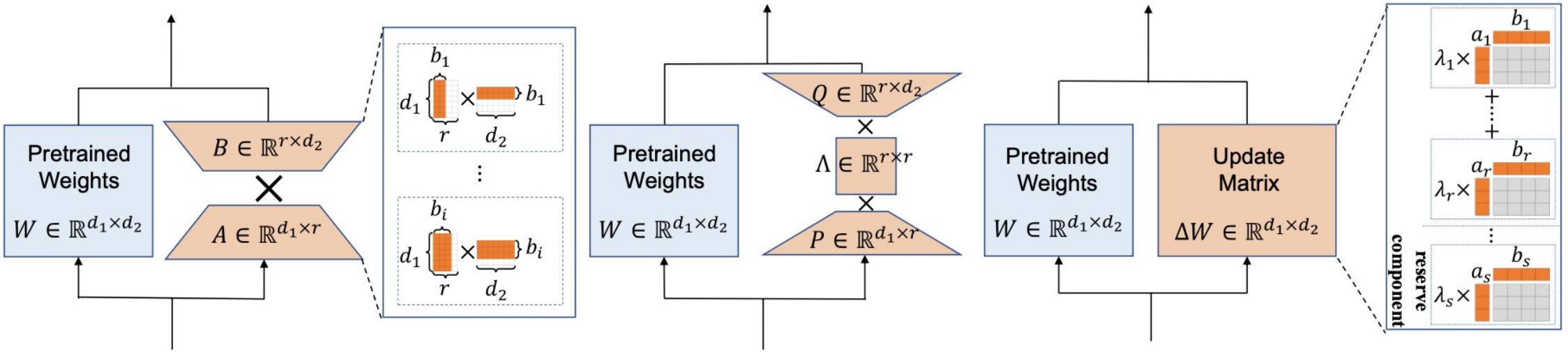
Why $A = \mathcal{N}(0, \sigma^2)$

Gradient updating process is stable

Implement for each layer:

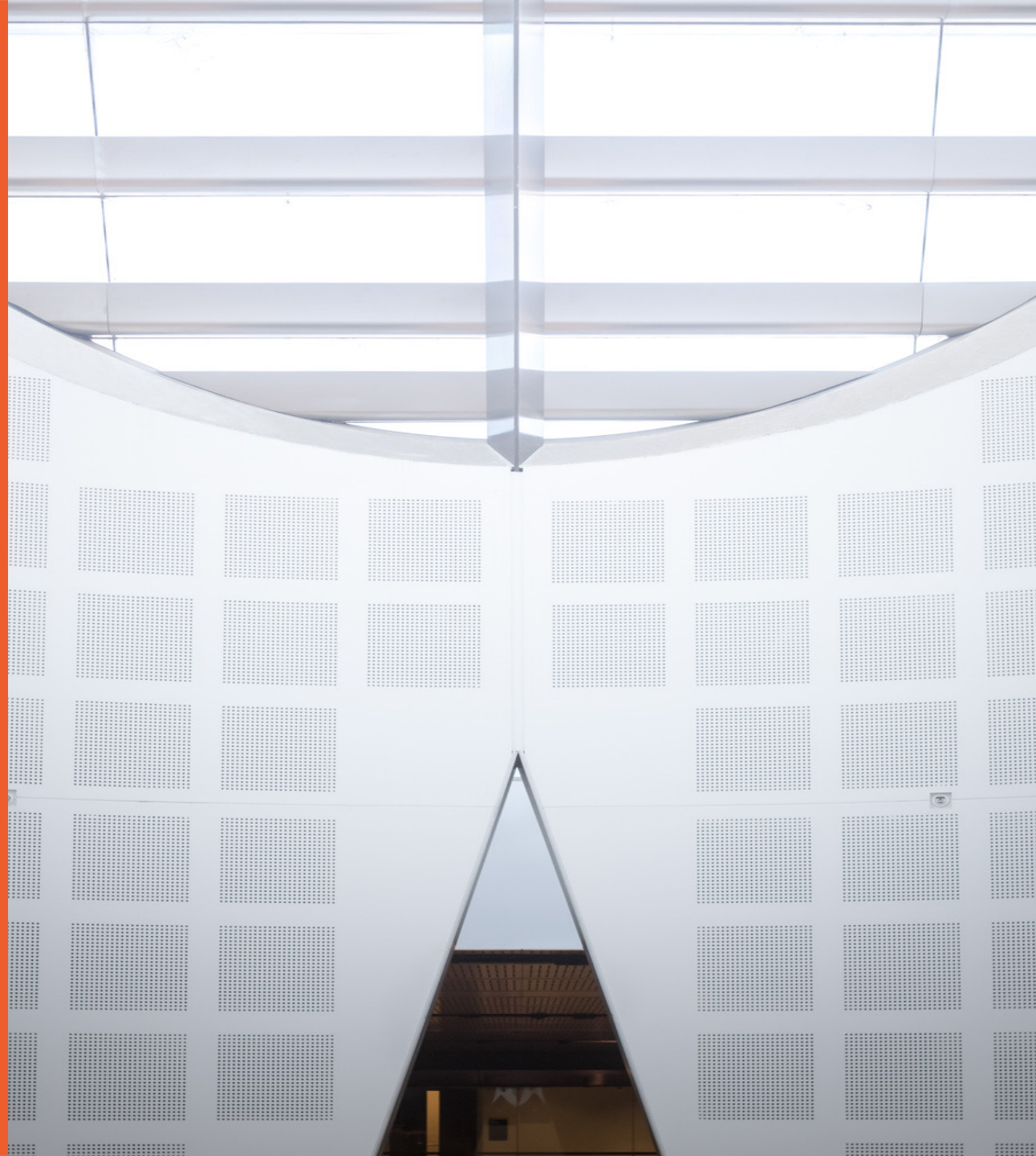


Other Variants



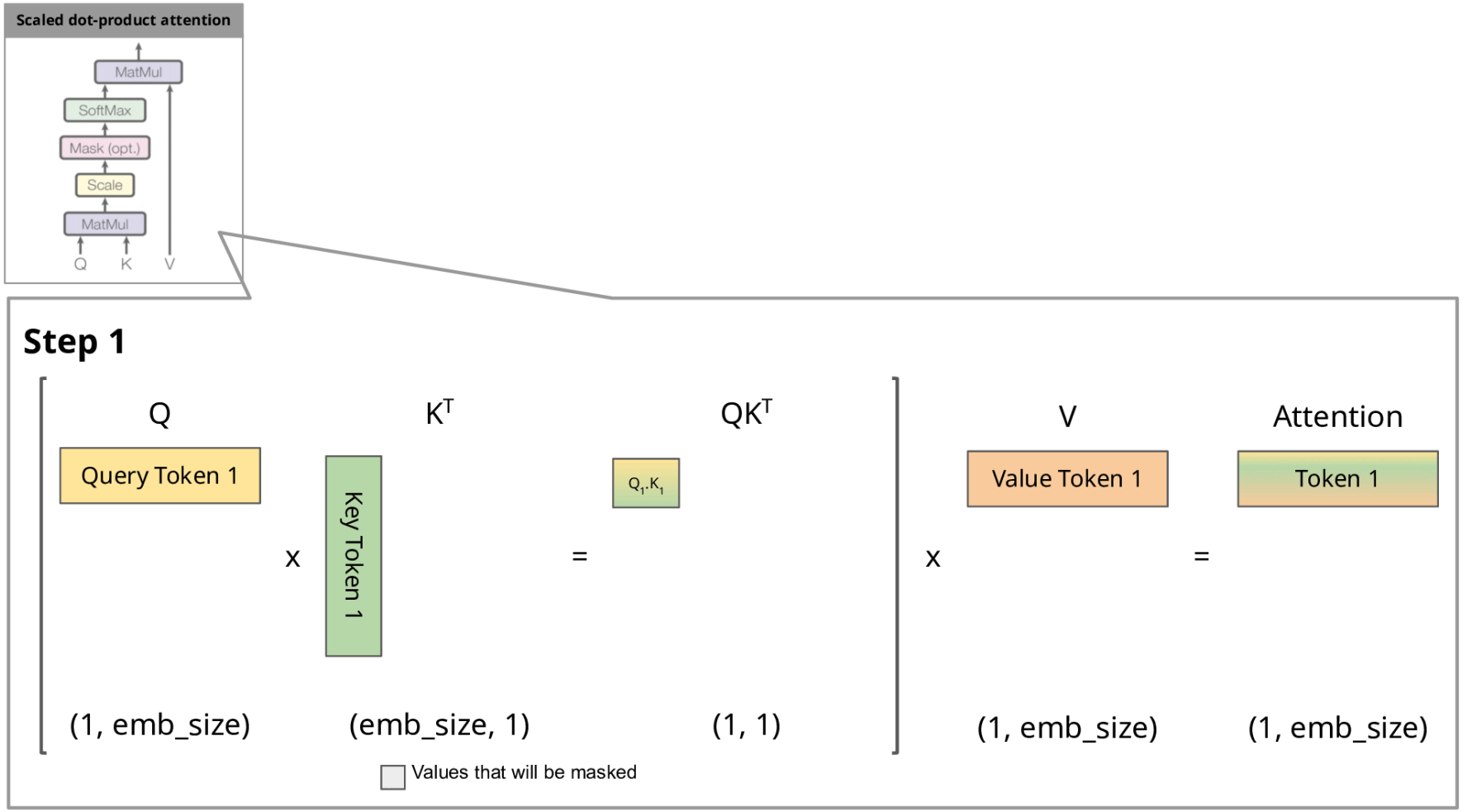
(a) DyLoRA (b) AdaLoRA (c) IncreLoRA

Seminar 1, Part 1.2: Efficiency of Inference



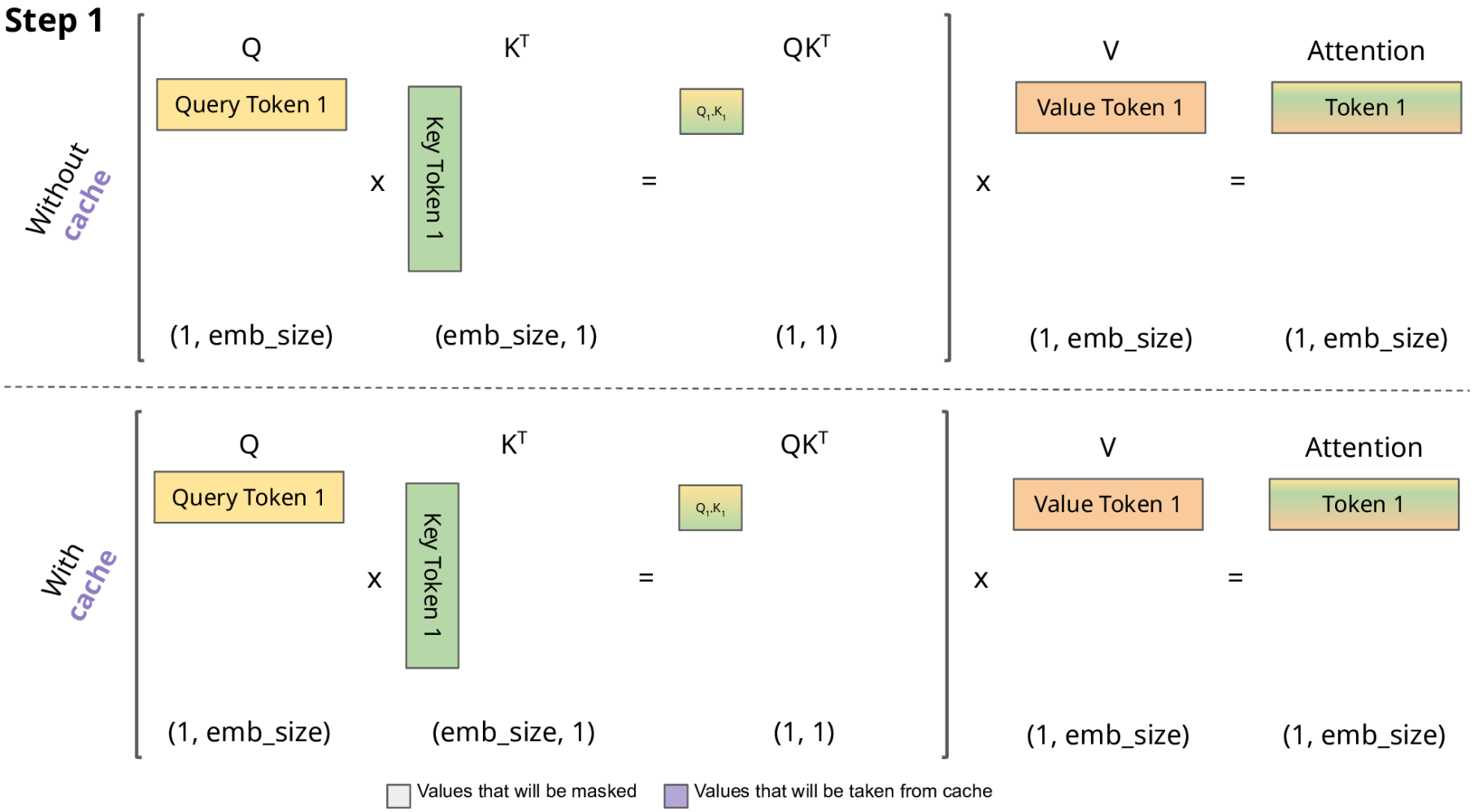
THE UNIVERSITY OF SYDNEY

KV Cache Inference

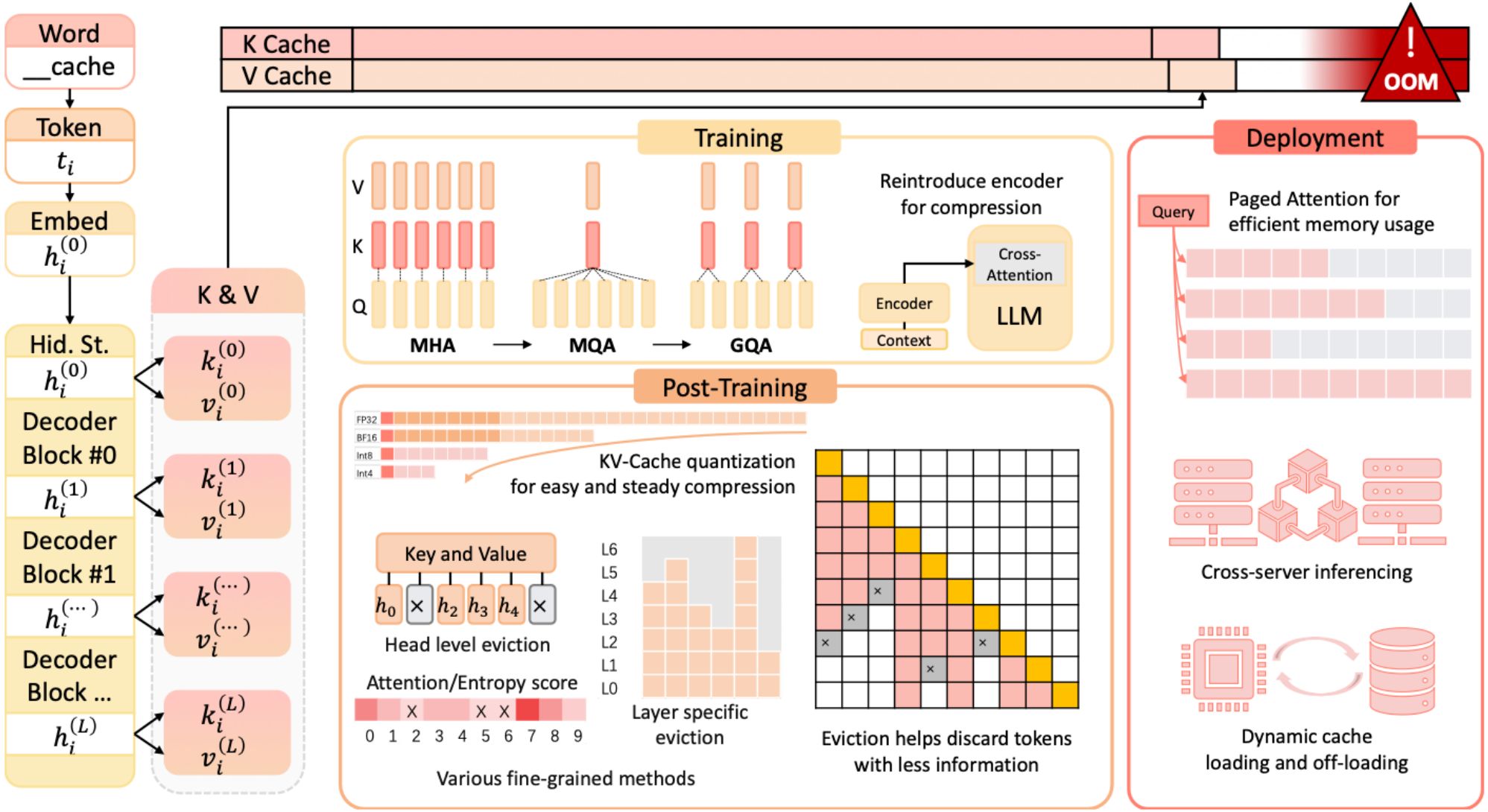


Zoom-in! (simplified without Scale and Softmax)

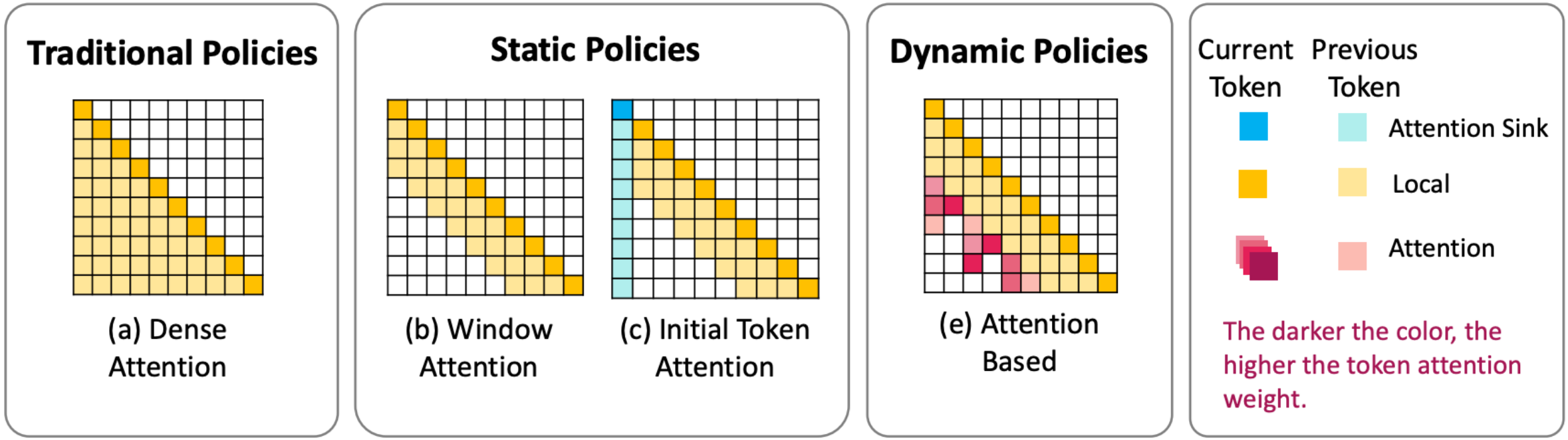
KV Cache Inference



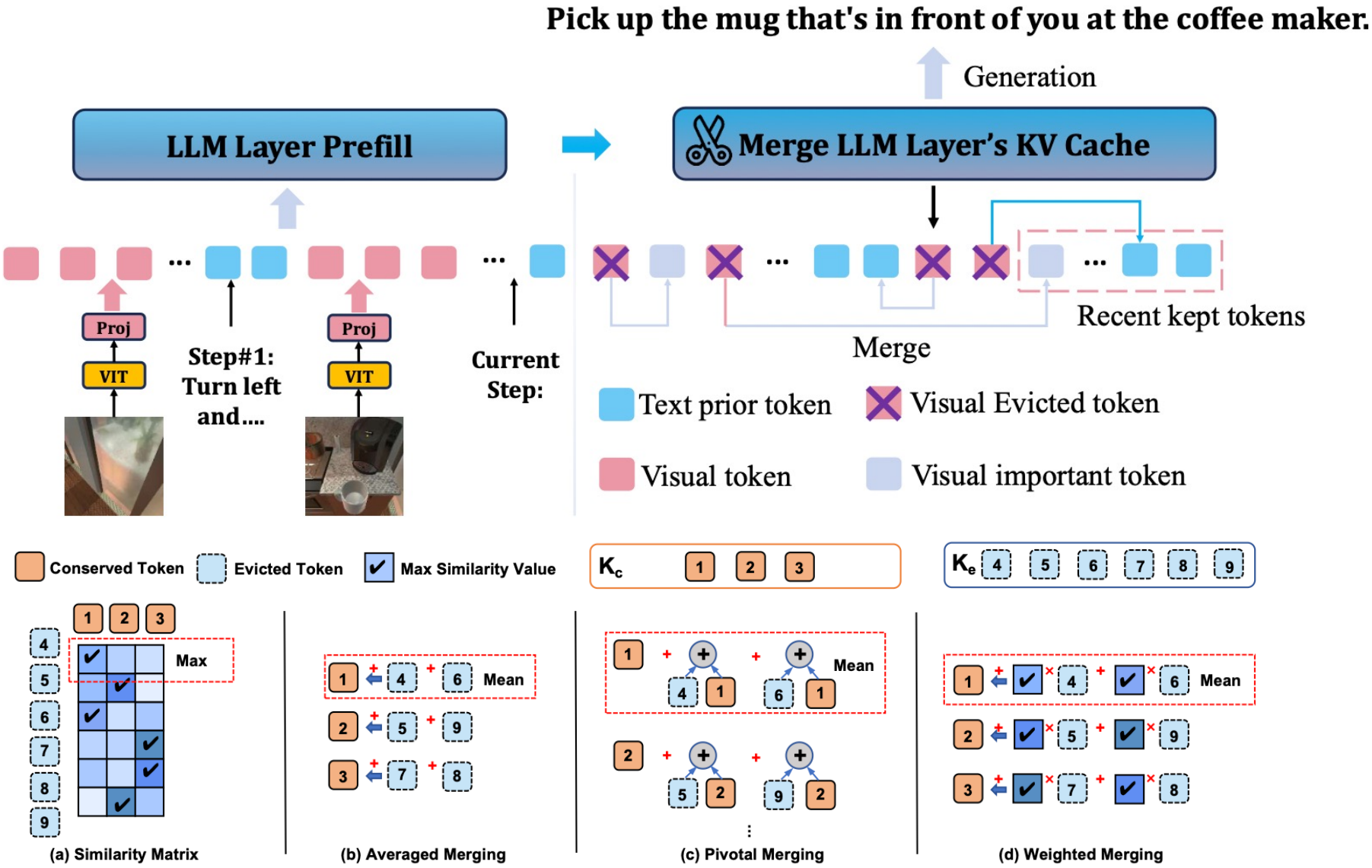
KV Cache Inference



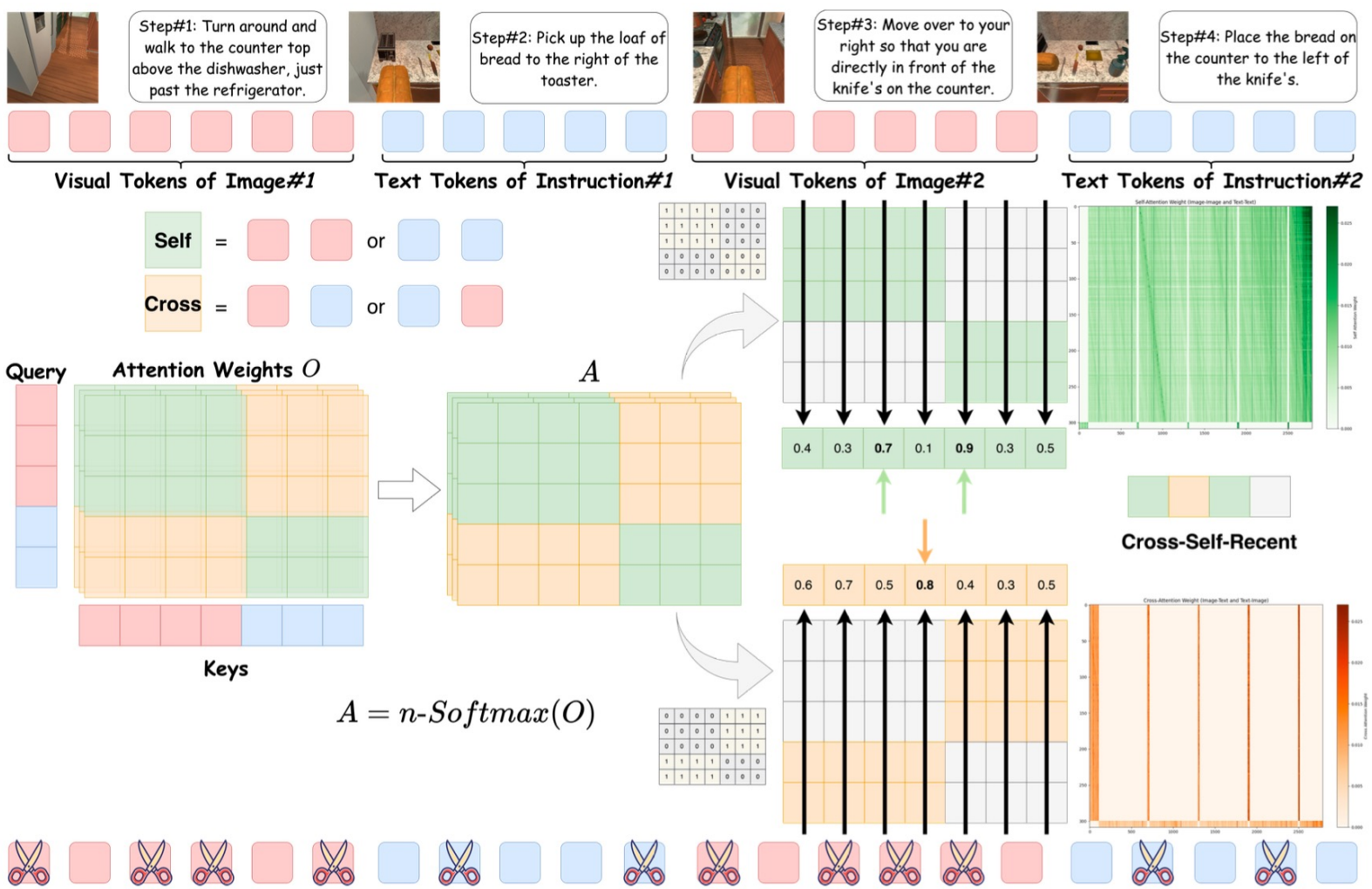
KV Cache Inference

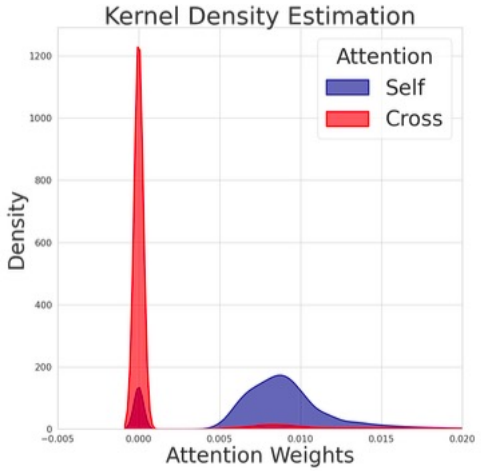


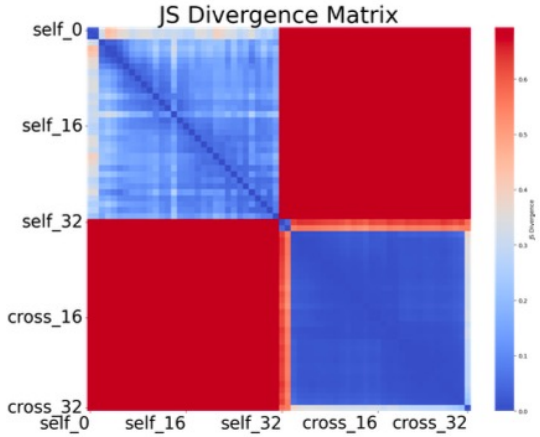
VLM Inference



VLM Inference







How about another approach ?

Thank you!
Q&A