Inference on Nano GPT - for Beginners

Pre-Requisites

Auto Regressive task ?

Time Vs Space Complexity

Inference and setting it up

Kv cacheing (although there is some issues in the code here)

Auto Regressive Task

The current output is depends upon the series of input which the model have seen through which can be neatly said in the equation given below

$$P(x_t|x_0,...,x_{t-1}) = \prod_{i=1}^{t} p(x_i|x_{i-1})$$

But in Language Modeling to pick a particular token Transformer does this by adding decoder casual mask

$$token = K(P(x_t|x_0,...,x_{t-1}))$$

$$\begin{cases} \text{argmax if } K = greedysampling \\ \text{Multinomial(topk) if } K = multinomalsampling \end{cases}$$

Time Vs Space complexity

- **Bubble sort: Time=O(n^2), Space=O(1)**
- **Merge sort: Time = O(nlogn) space O(logn)**
- **Heap Sort : Time= O(nlogn) space O(n)**

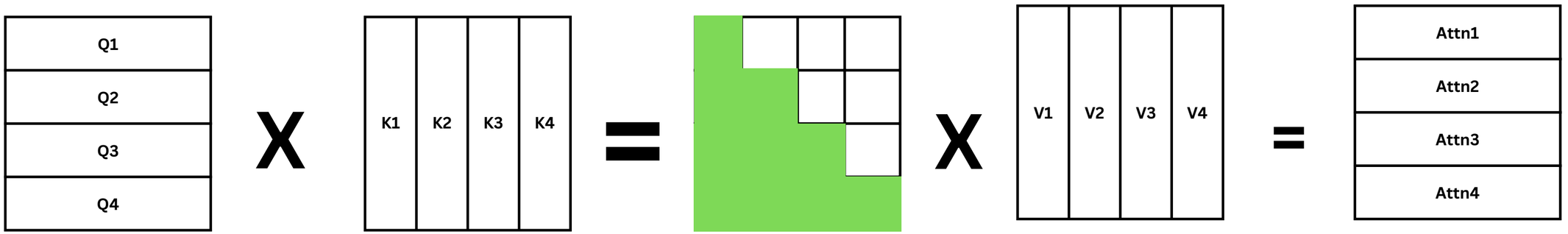
- We could observe that with Increasing space complexity we could kind of mitigate the time if correctly used.
- if space could be increased reasonable amount Time complexity could be kind of reduced

Whats that with transformers

Explanation of Forward pass of transformers if there is any confusion

Inference on Transformers

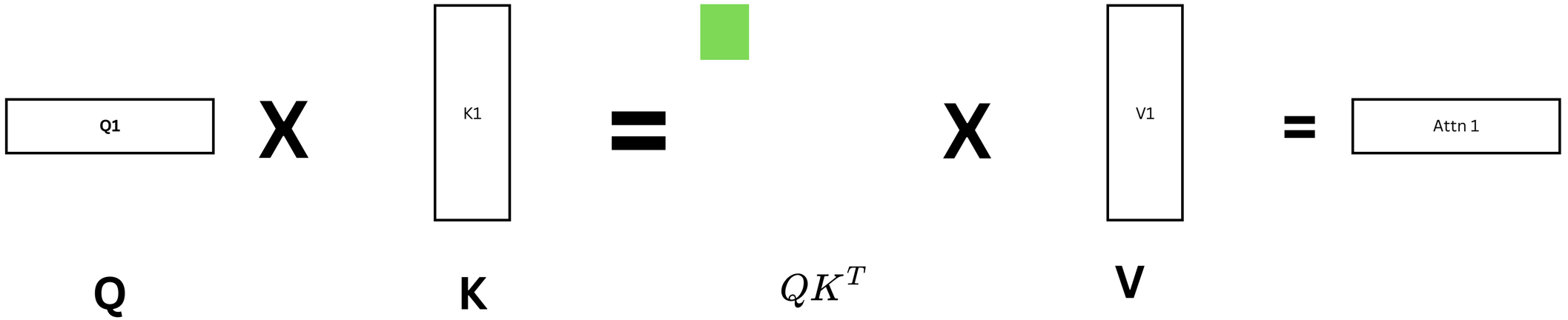
$$SM\left(\frac{QK^T}{\sqrt{d}}\right)V$$



Q **K** QK^T **V** **Attn**

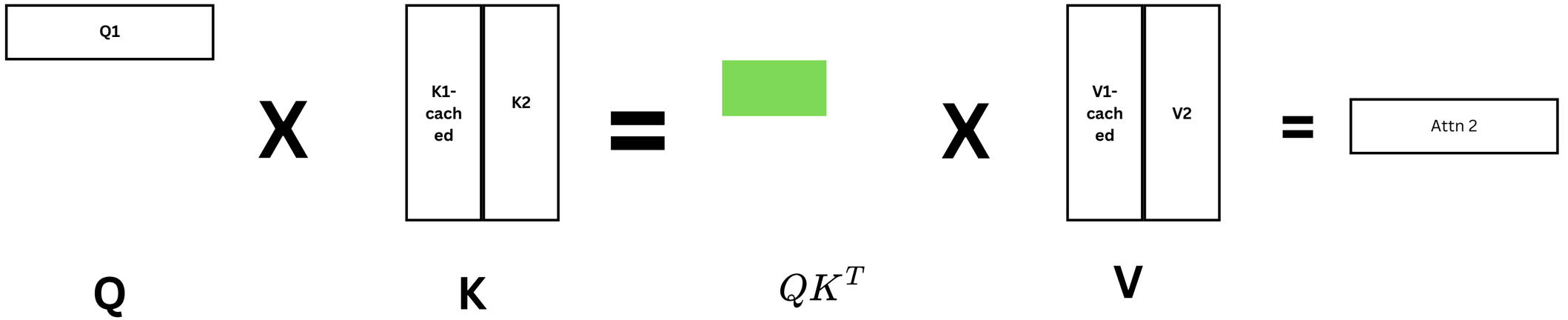
KV cacheing

step 1



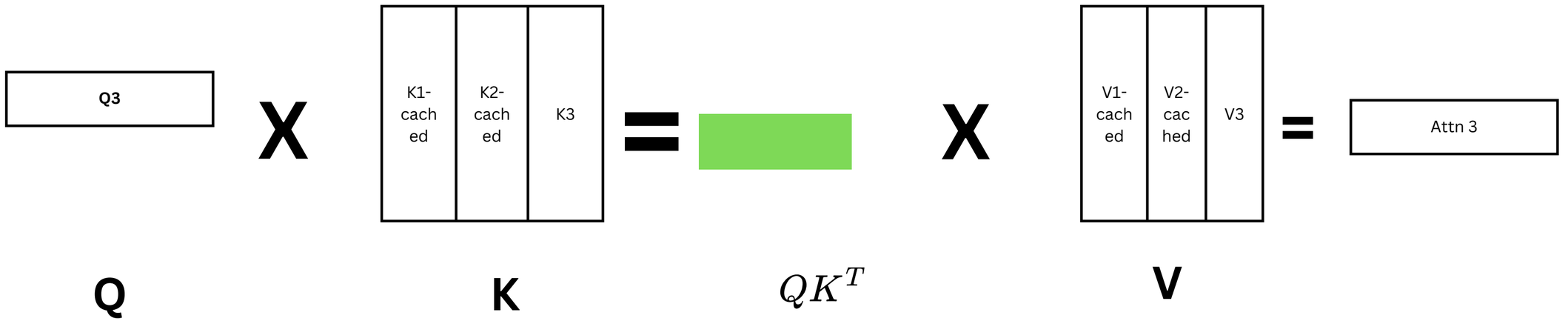
KV cacheing

step 1

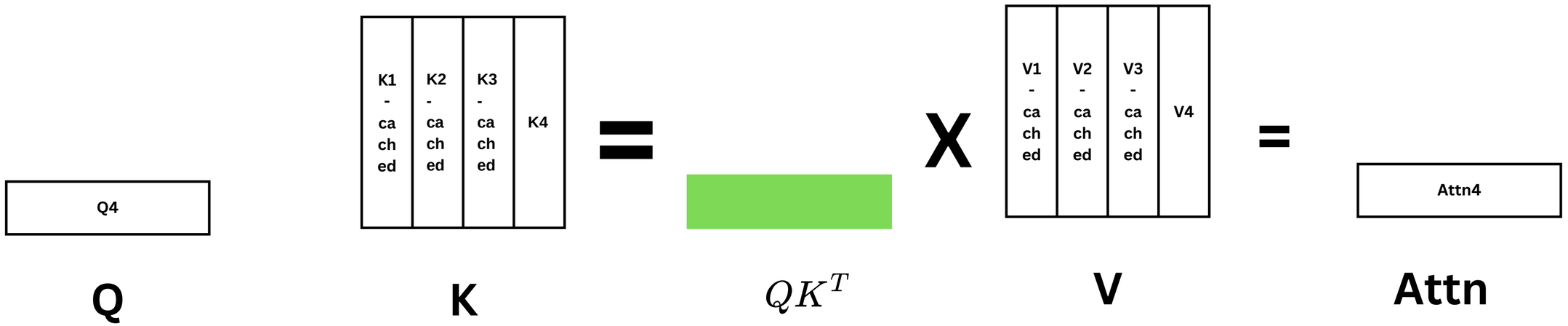


KV cacheing

step 1



KV cacheing


Q4
K1 - cached
K2 - cached
K3 - cached
K4
=
X
V1 - cached
V2 - cached
V3 - cached
V4
=
Attn4
Q
K
QK^T
V
Attn