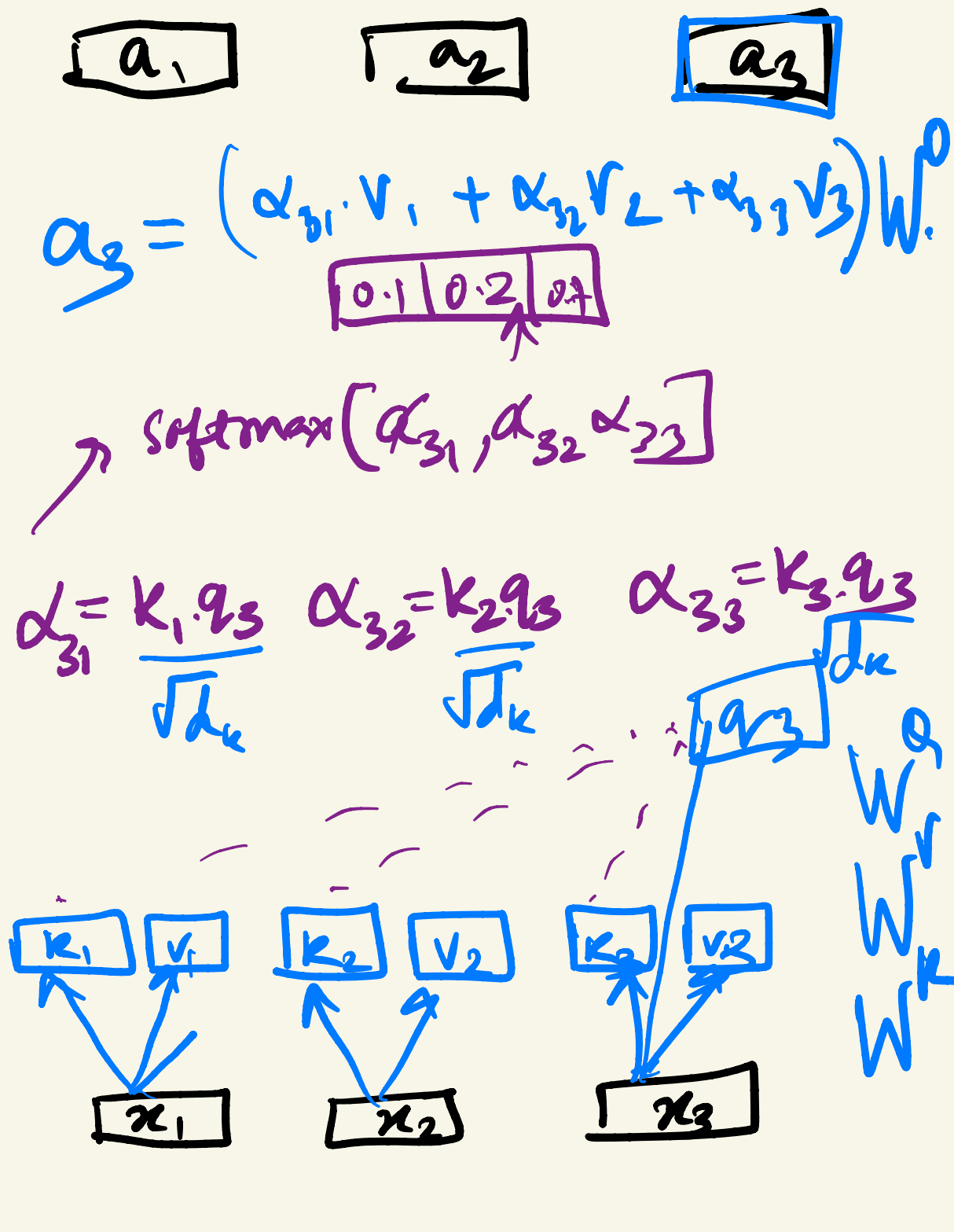
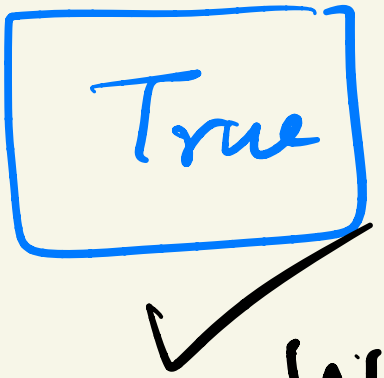

a1
a2
a3
a3 = (α31·v1 + α32v2 + α33v3) W^v
0.1
0.2
0.7
→ softmax[α31, α32 α33]
α31 = k1·q3 / √dk
α32 = k2·q3 / √dk
α33 = k3·q3 / √dk
q3
W^q
W^v
W^k
k1
v1
k2
v2
k3
v3
x1
x2
x3

The dimension of W^Q and W^K must be same.

True or False?

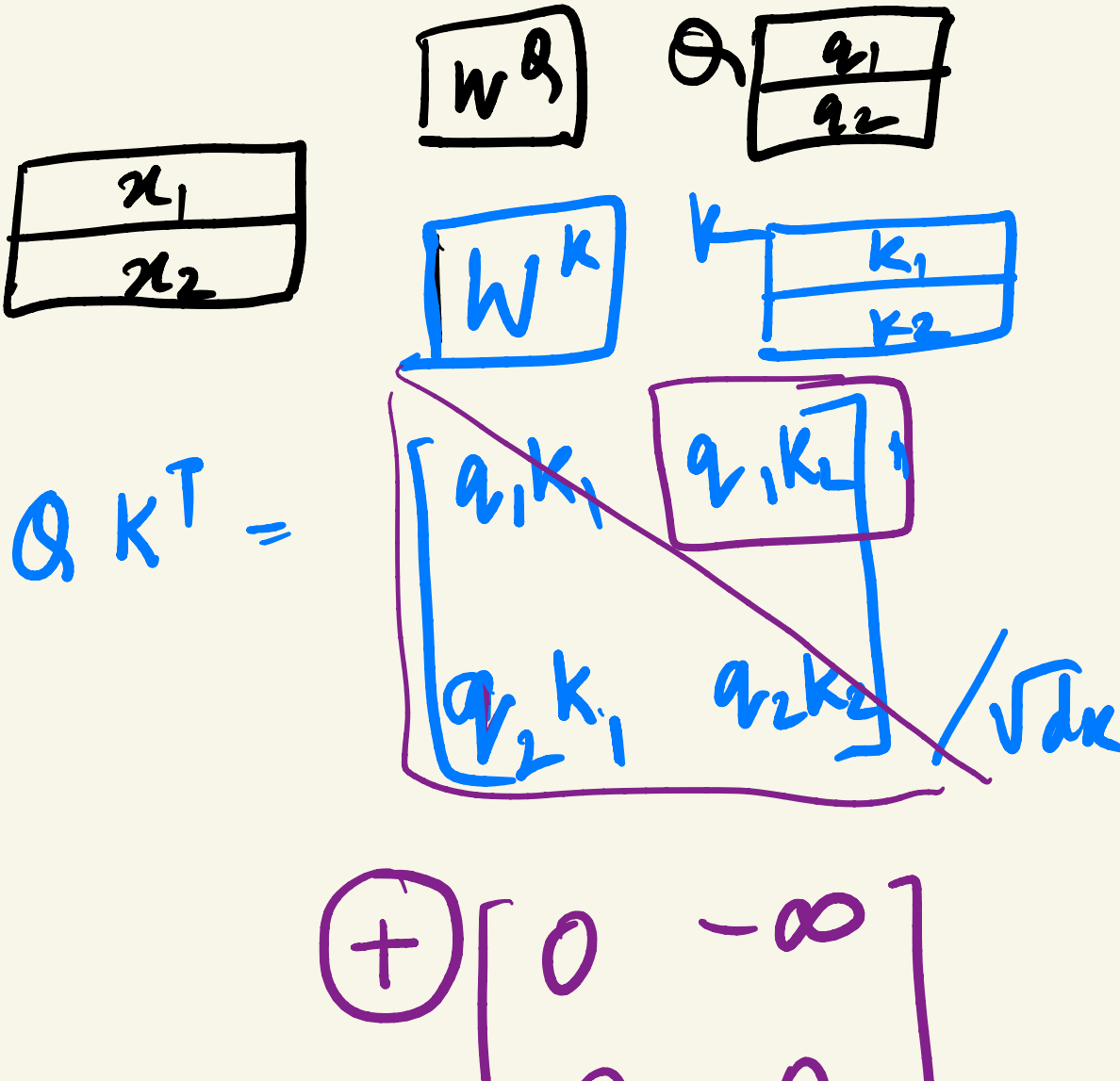


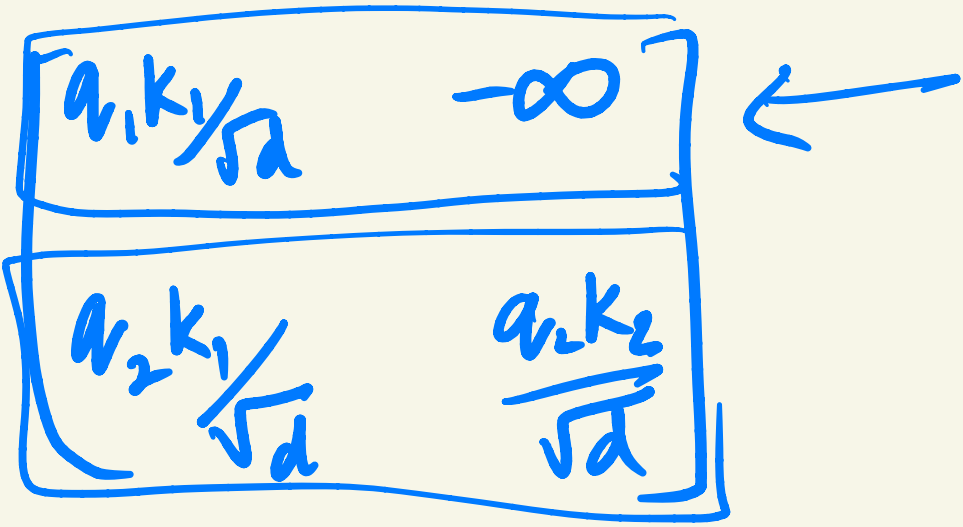
$$W^Q x_i = q_i$$
$$W^K x_j = k_j$$

The dimension of $W^{\vee}$ and W^{V} must be the same.

True or False?

In practice, usually the same.


W^Q
Q
q_1
q_2
x_1
x_2
W^K
K
k_1
k_2
$Q K^T =$
$q_1 k_1$
$q_1 k_2$
$q_2 k_1$
$q_2 k_2$
$/\sqrt{d_k}$
$+$
$\begin{bmatrix} 0 & -\infty \\ 0 & 0 \end{bmatrix}$
"causal mask"



After Matrix

$$\begin{bmatrix} 1 & 0 \\ \lambda & 1-\lambda \end{bmatrix} = H$$

$$(HV)W^{0} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}$$

Transformer model with query, key, value dim d, seq length n, no. of heads h.

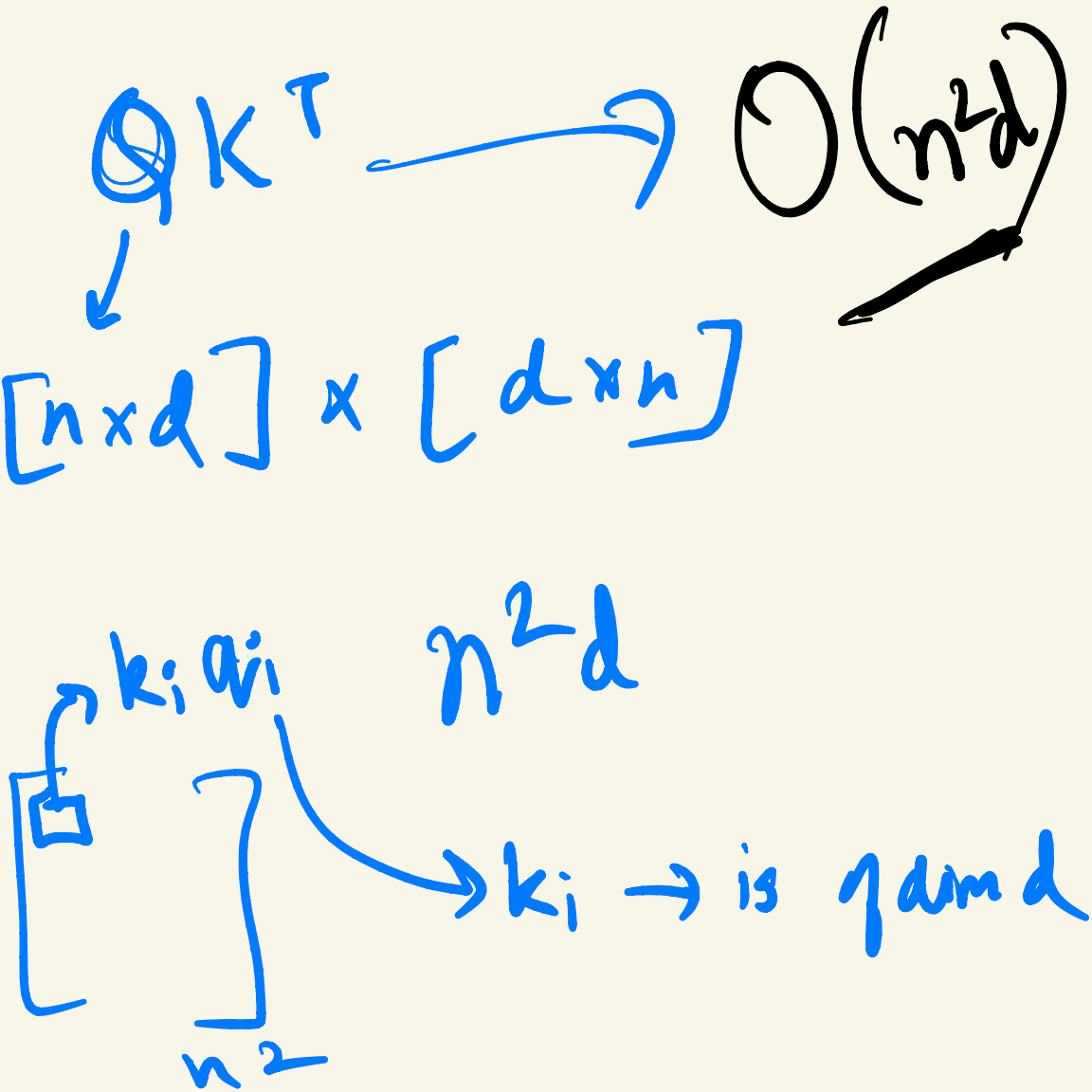
What is the complexity of self attention in 1 layer?

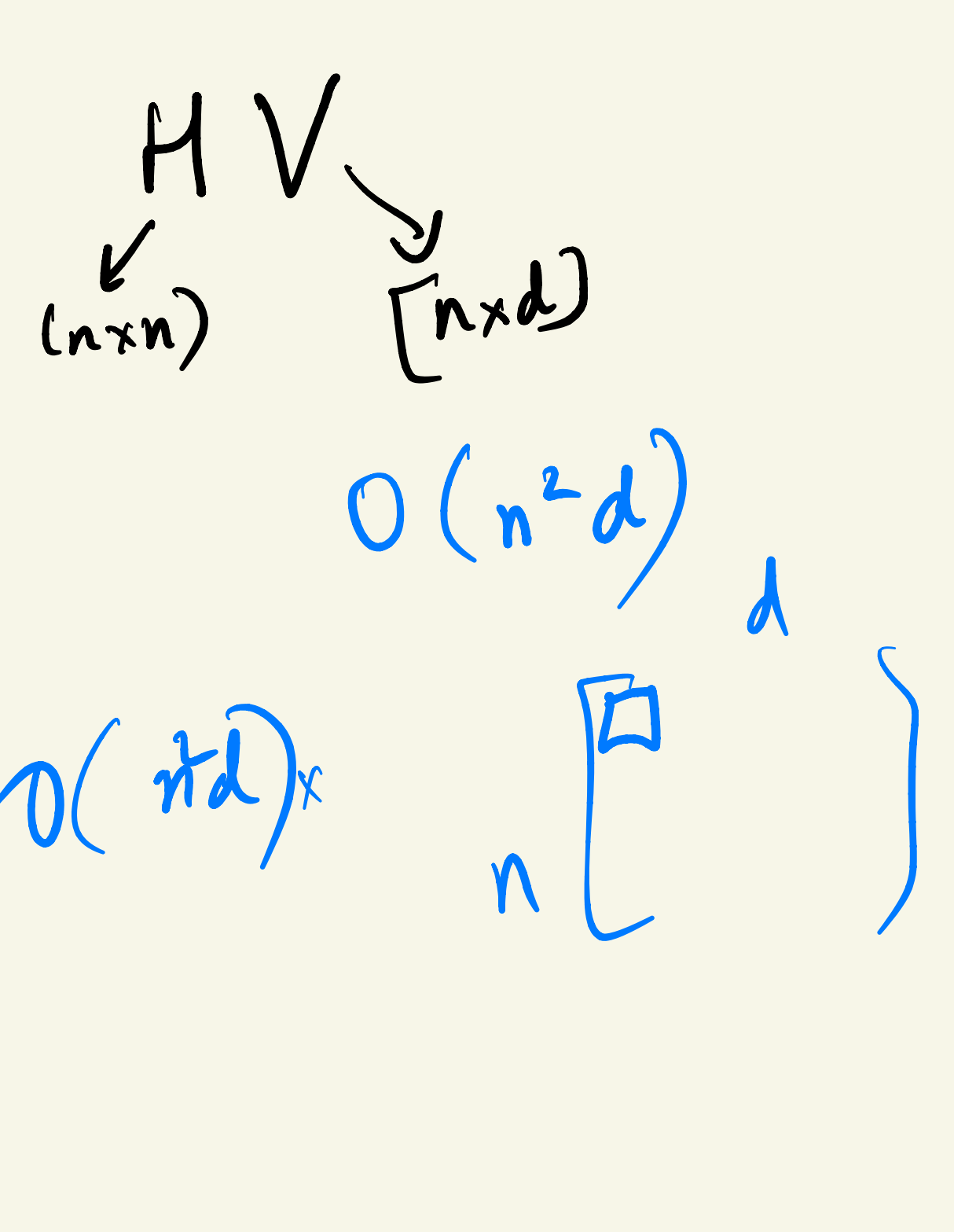
(A) $O(dnh)$

(B) $O(dn^2h)$

(C) $O(dn^2)$

(D) $O(d^2n^2)$


QK^T
O(n^2 d)
[n x d] x [d x n]
k_i q_i
n^2 d
k_i → is 1 dim d
n^2


H V
(n x n)
[n x d]
O(n²d)
d
O(n²d)x
n

$O(n^2 d h)$