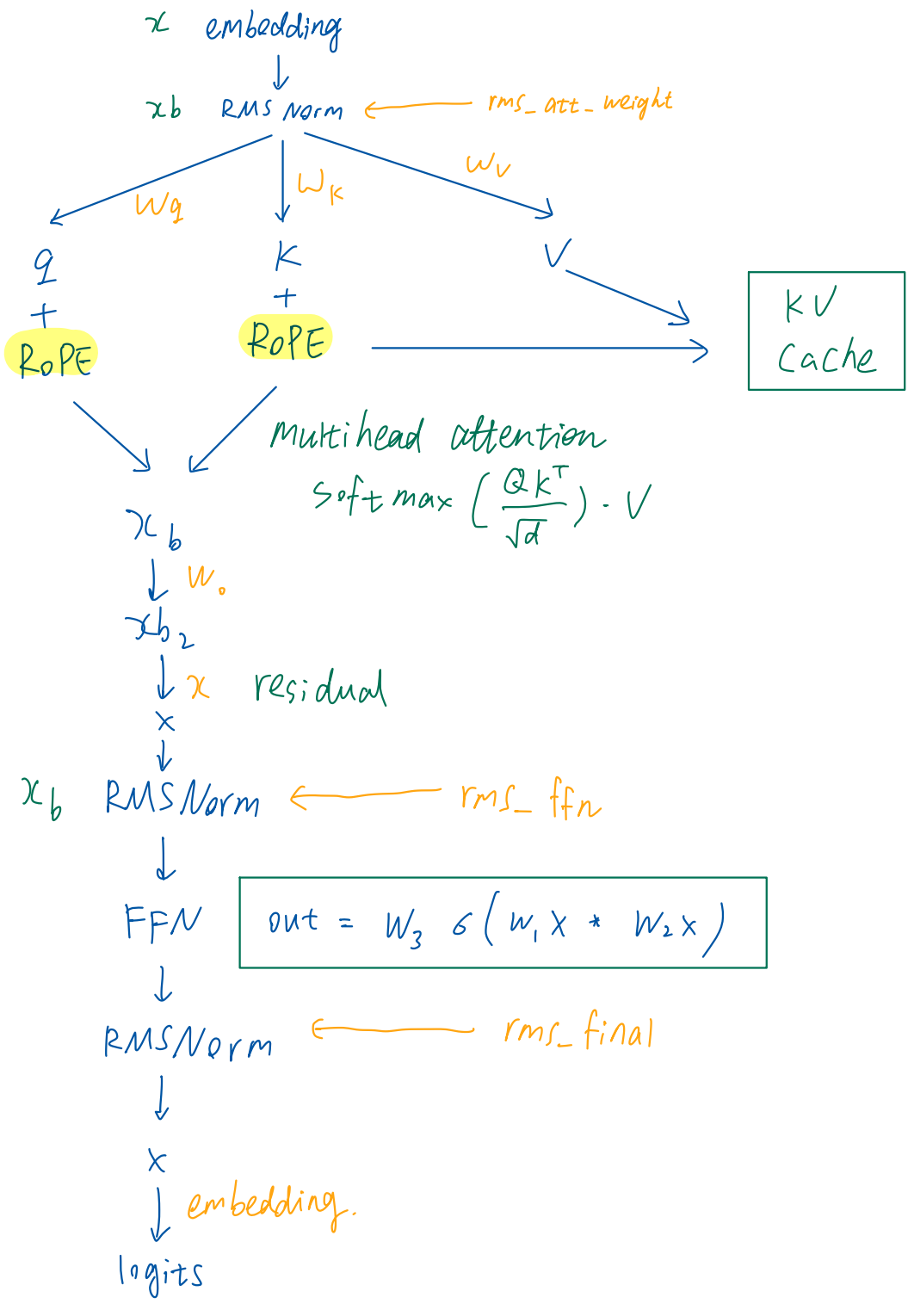

x embedding
xb RMS Norm
rms_att_weight
Wq
Wk
Wv
q
+
RoPE
k
+
RoPE
V
kV
Cache
multihead attention
Softmax $\left(\frac{Qk^T}{\sqrt{d}}\right) \cdot V$
x_b
W_o
x_{b_2}
x residual
x
x_b RMSNorm
rms_ffn
FFN
out = $W_3\, \sigma(W_1 x * W_2 x)$
RMSNorm
rms_final
x
embedding.
logits

RMS Norm:

$$y = \frac{x}{\sqrt{\frac{1}{N}\sum x^2 + \epsilon}} * \gamma$$

layer norm:

$$y = \frac{x - E(x)}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta$$

RoPE

$$\bar{q}_m = f_q(x_m, m) \qquad \bar{k}_n = f_k(x_n, n)$$

$$\langle f_q(x_m, m), f_k(x_n, n) \rangle = g(x_m, x_n, m-n)$$

$$f_q(x_m, m) = (W_q x_m) e^{im\theta} \qquad f_k(x_n, n) = (W_k x_n) e^{in\theta}$$

$$g(x_m, x_n, m-n) = \mathrm{Re}\left[(W_q x_m)(W_k x_n)^* e^{i(m-n)\theta}\right]$$

$$\begin{bmatrix} x_m^{i'} & x_m^{j'} \end{bmatrix} = \begin{bmatrix} x_m^i & x_m^j \end{bmatrix} \begin{bmatrix} \cos(m\theta_m) & -\sin(m\theta_m) \\ \sin(m\theta_m) & \cos(m\theta_m) \end{bmatrix}$$

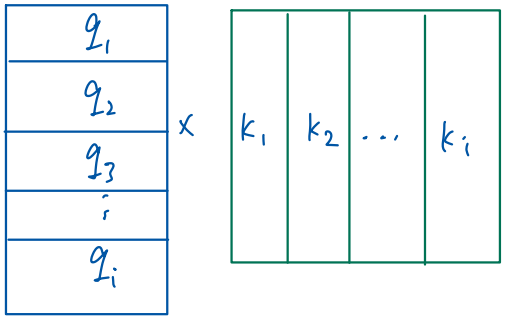
KV Cache

Q: (seq len, dim)

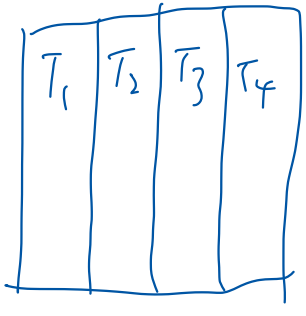
K: (dim, seq len)

V: (seq len, dim)

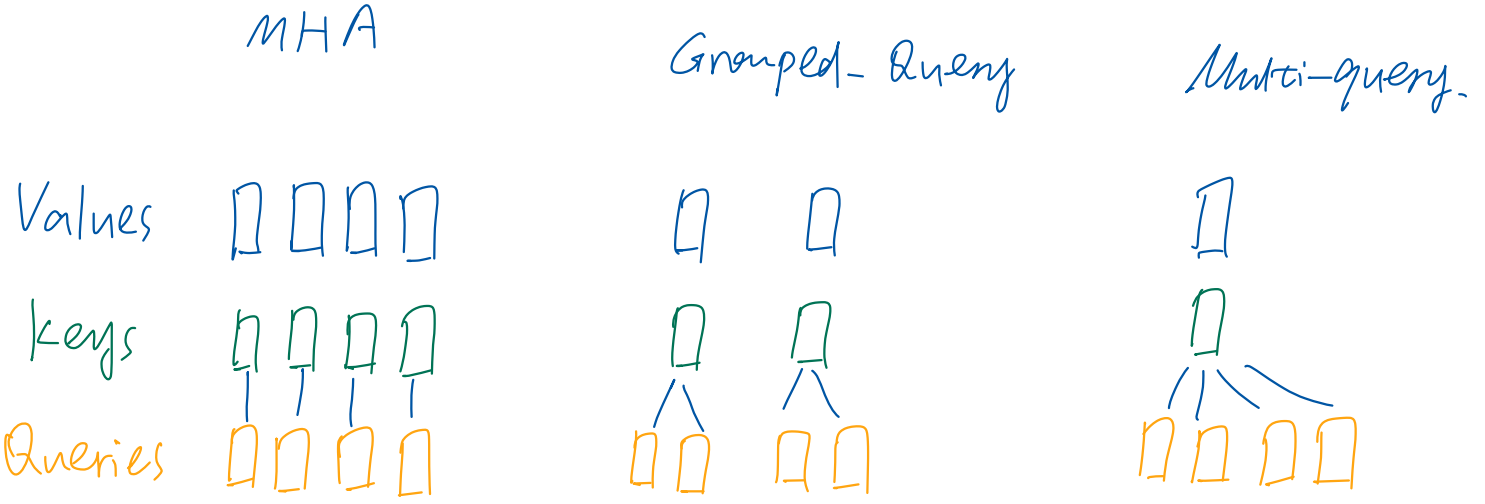
Z: (seq. len, dim)



Q $\quad$ K^T $\quad$ QK^T



Grouped Multi-Query Attention.



Feedforward

$$out = W_3\, \sigma(W_1 x * W_2 x)$$

$$SiLU(x) = \frac{x}{1 + e^{-x}}$$

$$hb = x_b^T W_1$$

$$hb_2 = x_b^T W_3$$

$$hb = SiLU(hb)$$

$$hb = hb * hb2$$

$$xb = hb^T W_2$$

$$x = x_b + x$$

Transformer:

$$out = \sigma(W_1 x) W_2$$