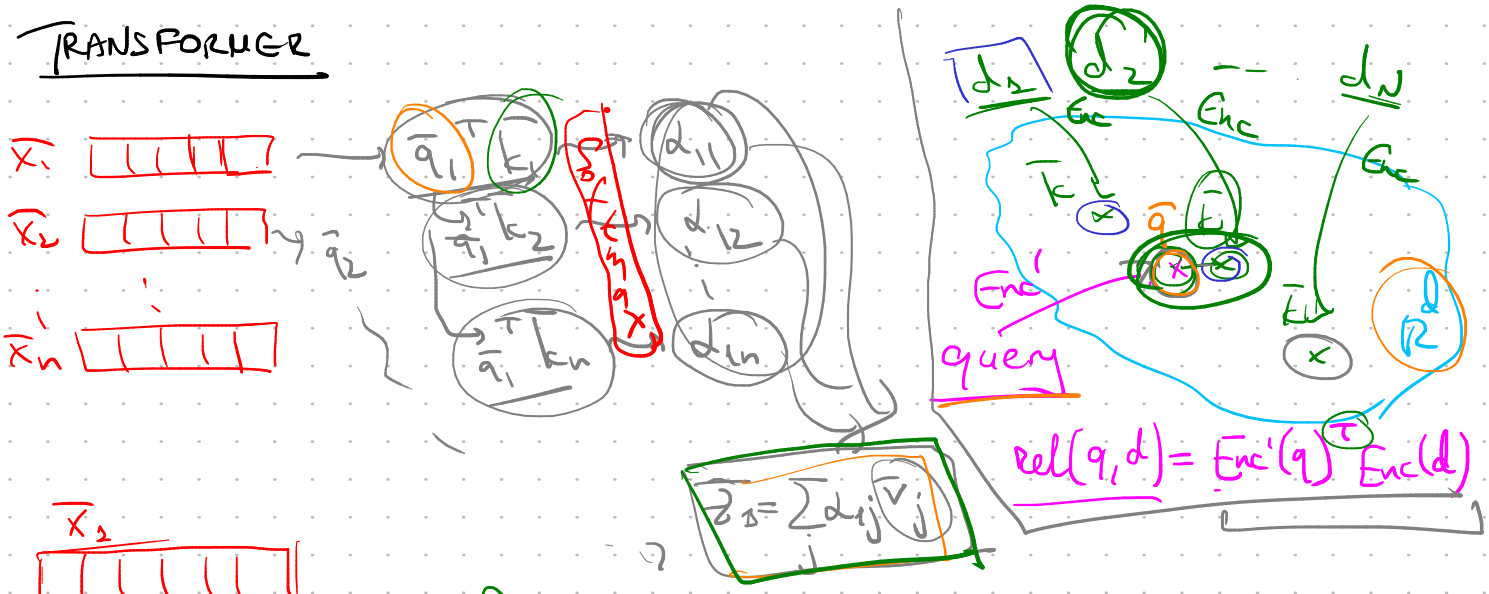
TRANSFORMER

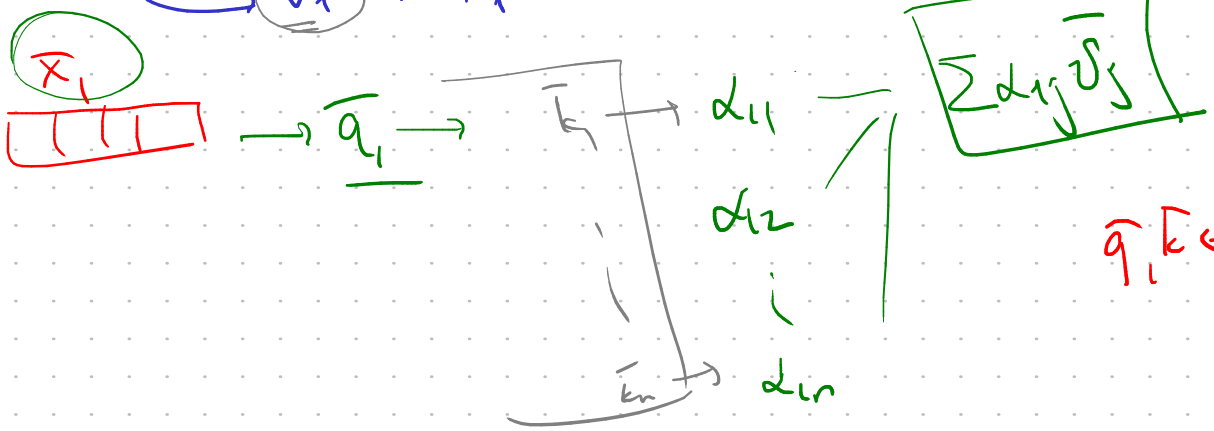


$\bar{x}_1$

$\bar{q}_1 = W^Q \bar{x}_1$

$\bar{k}_1 = W^K \bar{x}_1$

$\bar{v}_1 = W^V \bar{x}_1$



$\bar{q}_1, \bar{k} \in \mathbb{R}^d$

$\bar{x}_1 \rightarrow \bar{q}_1 \rightarrow \frac{1}{\sqrt{d}} \bar{q}_1^T \bar{k}_1 \quad \frac{1}{\sqrt{d}} \bar{q}_1^T \bar{k}_2 \quad \cdots \quad \frac{1}{\sqrt{d}} \bar{q}_1^T \bar{k}_n$

$\downarrow \qquad \downarrow \qquad \qquad \downarrow$

softmax

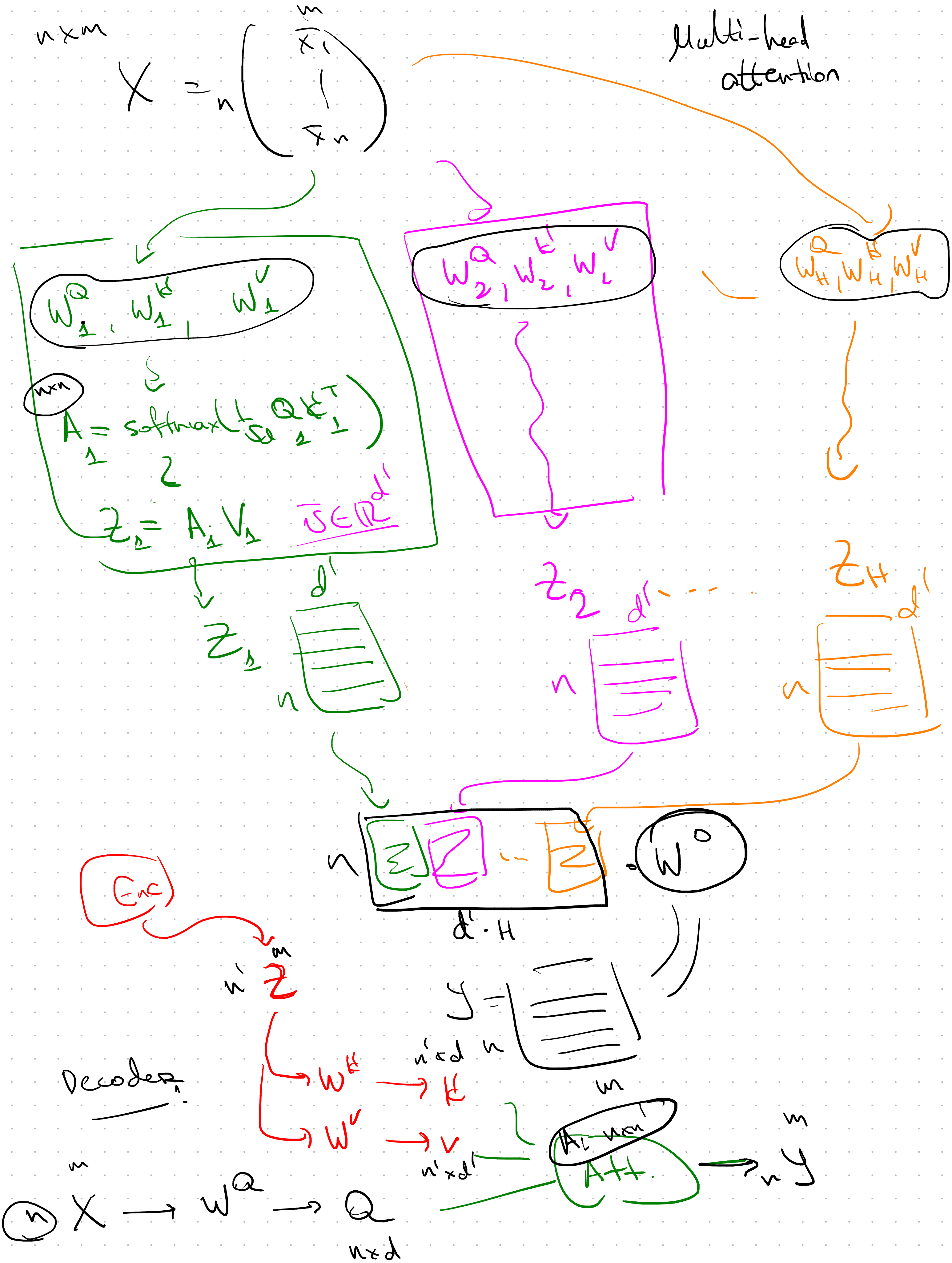
$\uparrow \qquad | \qquad |$

$$\alpha_{1j} = \frac{e^{1/\sqrt{d} \, \bar{q}_1^T \bar{k}_j}}{\sum_s e^{1/\sqrt{d} \, \bar{q}_1^T \bar{k}_s}}$$

$$\bar{z}_1 = \sum_j \alpha_{1j} \bar{v}_j = \sum_{j=1}^{n} \text{softmax}\left(\frac{1}{\sqrt{d}} \bar{q}_1^T \bar{k}_j\right) \bar{v}_j$$

$$Z = \text{softmax}\left(\frac{1}{\sqrt{d}} Q K^T\right) \cdot V$$

где $Q = W^Q X$, $K = W^K X$, $V = W^V X$



Multi-head attention
n×m
X = (x_1 … x_n)
W_1^Q, W_1^K, W_1^V
n×n
A_1 = softmax(1/√d Q_1 K_1^T)
Z_1 = A_1 V_1
V ∈ ℝ^{d'}
W_2^Q, W_2^K, W_2^V
W_H^Q, W_H^K, W_H^V
Z_1
Z_2
Z_H
d'
n
W^O
d'·H
Y =
Enc
Z
W^K
W^V
K
V
n'×d
n'×d'
Decoder:
X → W^Q → Q
n×d
A_i n×n'
Att.
Y