*Attention:*

**Encoder**

**Decoder**

**Encoder Hidden States**

```
< y_1 >
Ich
< y_2 >
else
< y_3 >
ein
< y_4 >
Apfel
```

```
This hidden state encodes the entire input sequence <s> and is used to "start off" the decoder.
```

```
< a_1 >
I
< a_2 >
eat
< a_3 >
an
< a_4 >
apple
```
Construct Keys & Values from Encoder Hidden States

Using matrices $W_k$ and $W_v$, we construct keys & values for all the Encoder Hidden States.

Matrix Dimensions:

$W_k = h_{hid} \times h_k$

$W_v = h_{hid} \times h_v$
"Construct Queries from Decoder Hidden States"

Using matrix $W_q$, we construct a query from the Decoder hidden state.

$W_q = h_{wq} \times h_k$

* NOTE: Queries & Keys are of the same dimension in this example but this is not true for all types of attention.
"Compute Attention Scores"

\[
\text{Softmax} \left( \begin{pmatrix} h_k \\ z_c \end{pmatrix} \right) = \begin{pmatrix} \vdots \\ \vdots \end{pmatrix}
\]

"Scale values by attention scores"

\[
\begin{pmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{pmatrix} \times \begin{pmatrix} \vdots \\ \vdots \end{pmatrix} \times \begin{pmatrix} \vdots \\ \vdots \end{pmatrix} \times \begin{pmatrix} \vdots \\ \vdots \end{pmatrix}
\]

\[
\text{Sum} \rightarrow h_v
\]

"Final Representation for One Decoder State"
The Complete Picture:

Attention context for the first time step, compute context for $y_0$ using Attention
- Similarly, repeat for all consecutive time steps.
- Notice that Decoder state 1 depends on the input from state 0. Therefore, there is a left-to-right dependence.
Keys, values, and queries are all computed from encoder hidden states. BUT, no left-to-right dependence, so can compute all context vectors in parallel.
"Masking in Transformer Decoder"

- want to condition only on previous tokens' context. "Mask" out the keys from consecutive tokens while computing attention.

\[
\text{Softmax}
\begin{pmatrix}
\text{h}_i \\
\text{q}_i
\end{pmatrix}
\times \text{Mask}_i
\]

- Mask for \( q_0 = [1 \ 0 \ 0 \ 0] \)
- Mask for \( q_1 = [1 \ 1 \ 0 \ 0] \)
- Mask for \( q_2 = [1 \ 1 \ 1 \ 0] \)
- Mask for \( q_3 = [1 \ 1 \ 1 \ 1] \)

* To compute all the attention scores at once:

\[
\text{Softmax}
\begin{pmatrix}
\text{Query Matrix} \\
T \times h_k
\end{pmatrix}
\times
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1
\end{pmatrix}
\]