Sequence Models and Attention

CS280

Spring 2025

Angjoo Kanazawa

Next few lectures

- Today: Transformers / Attention
- Next: Vision Transformers, DINO
- Next Week: Diffusion Models

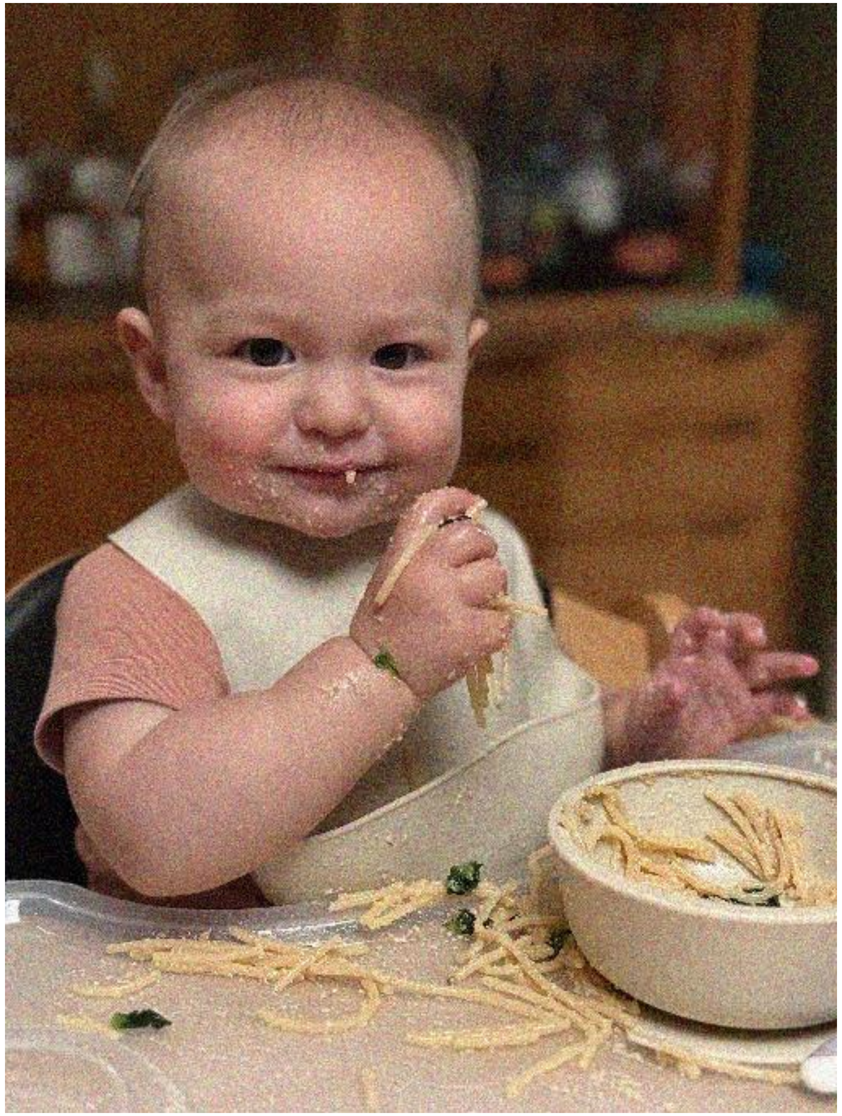
What is Attention

A Computer Vision perspective

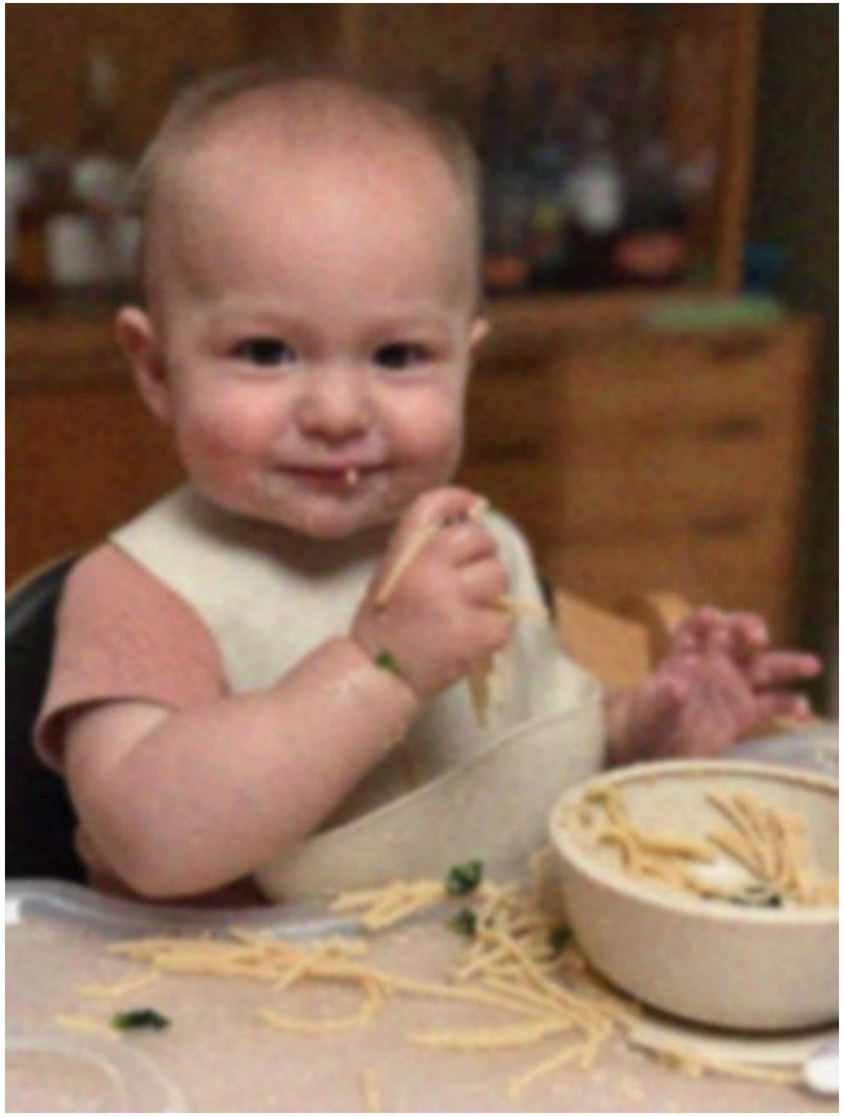
Computer Vision History on filtering

- Gaussian Filter
- Anisotropic Diffusion
- Bilateral Filter
- Non-Local Means

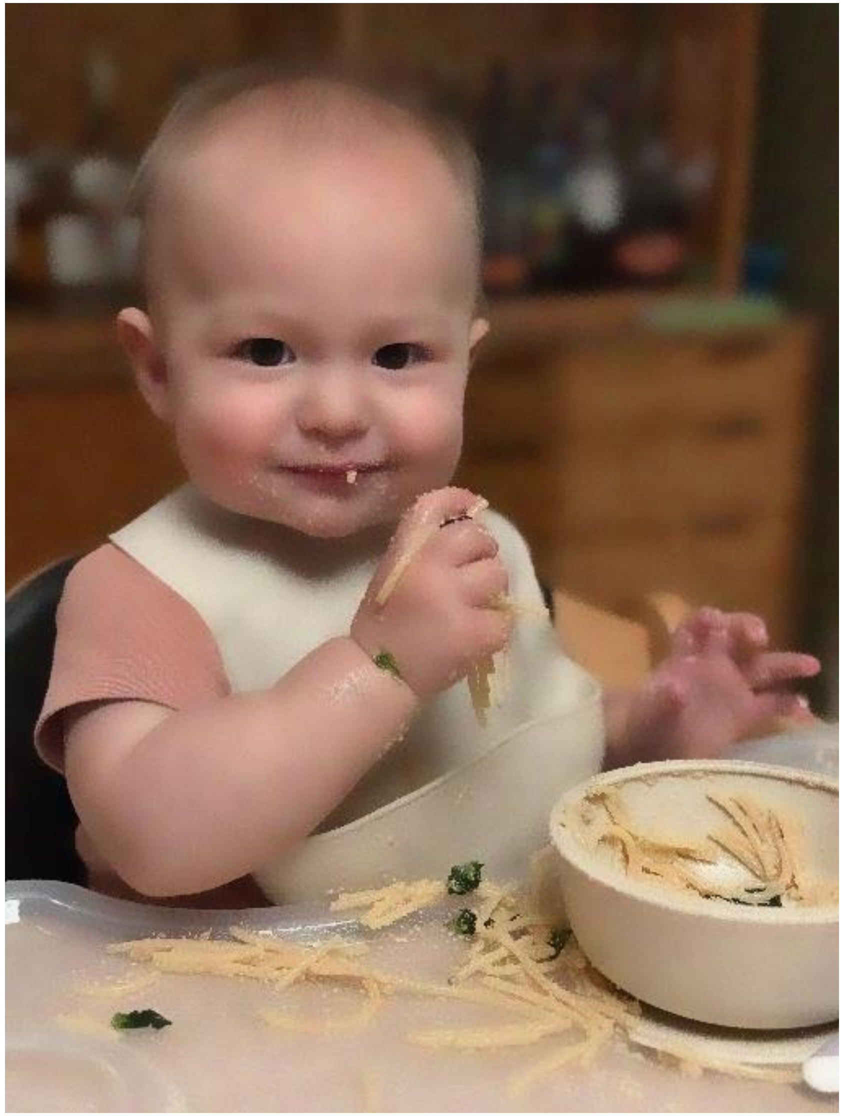
How to filter an image?



Gaussian filter



How to do edge preserving filter?



Physical analog: Heat diffusion

Heat Equation

$$\frac{\partial u}{\partial t} = c\Delta u = \mathrm{div}(c\nabla u)$$

Moving from higher to lower concentration of signal u

Physical analog: Heat diffusion

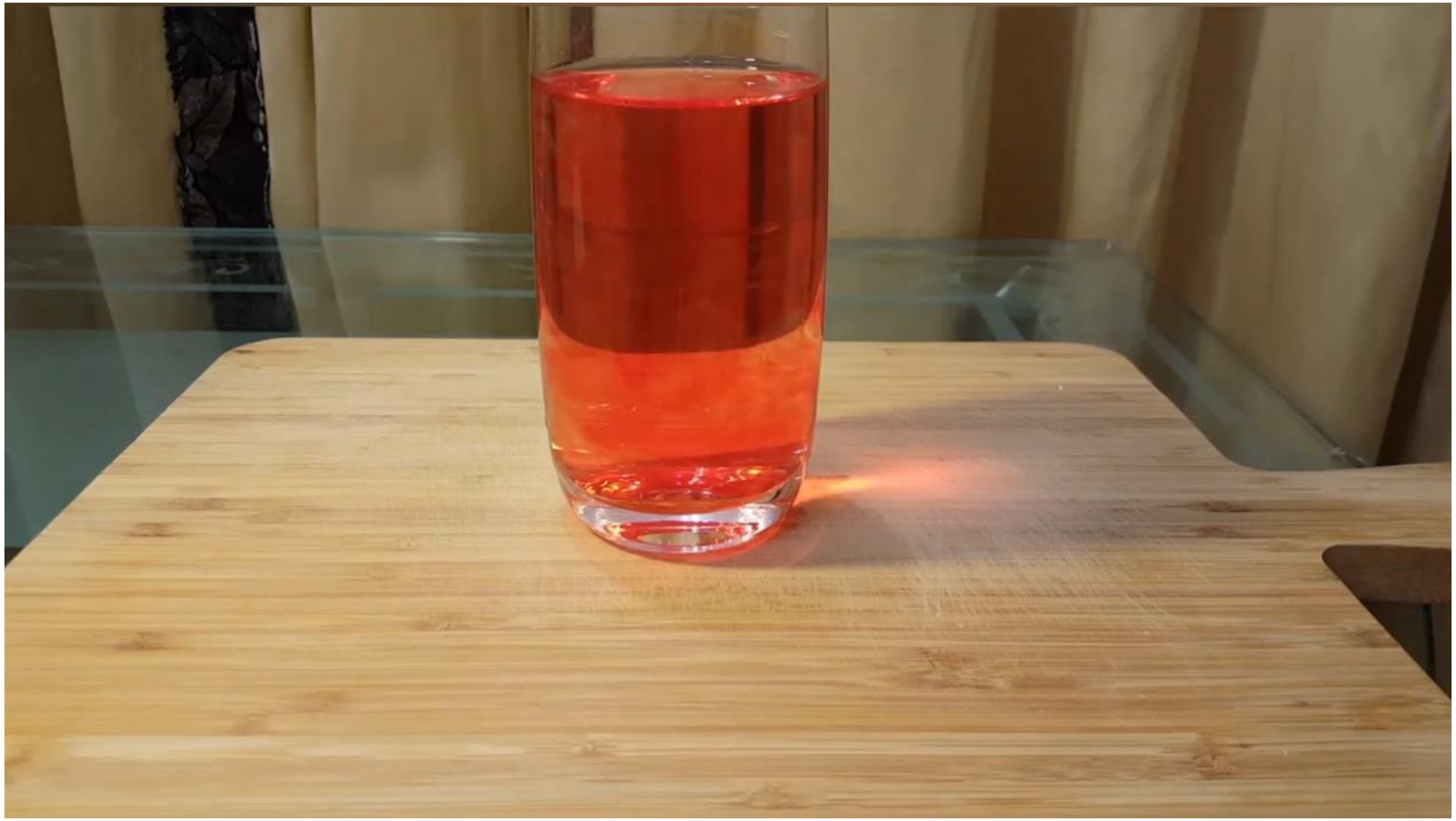
Heat Equation

$$\frac{\partial u}{\partial t} = c\Delta u = \mathrm{div}(c\nabla u)$$

Moving from higher to lower concentration of signal u

Solution is exactly convolution with a gaussian kernel!

Useful intuition for later…



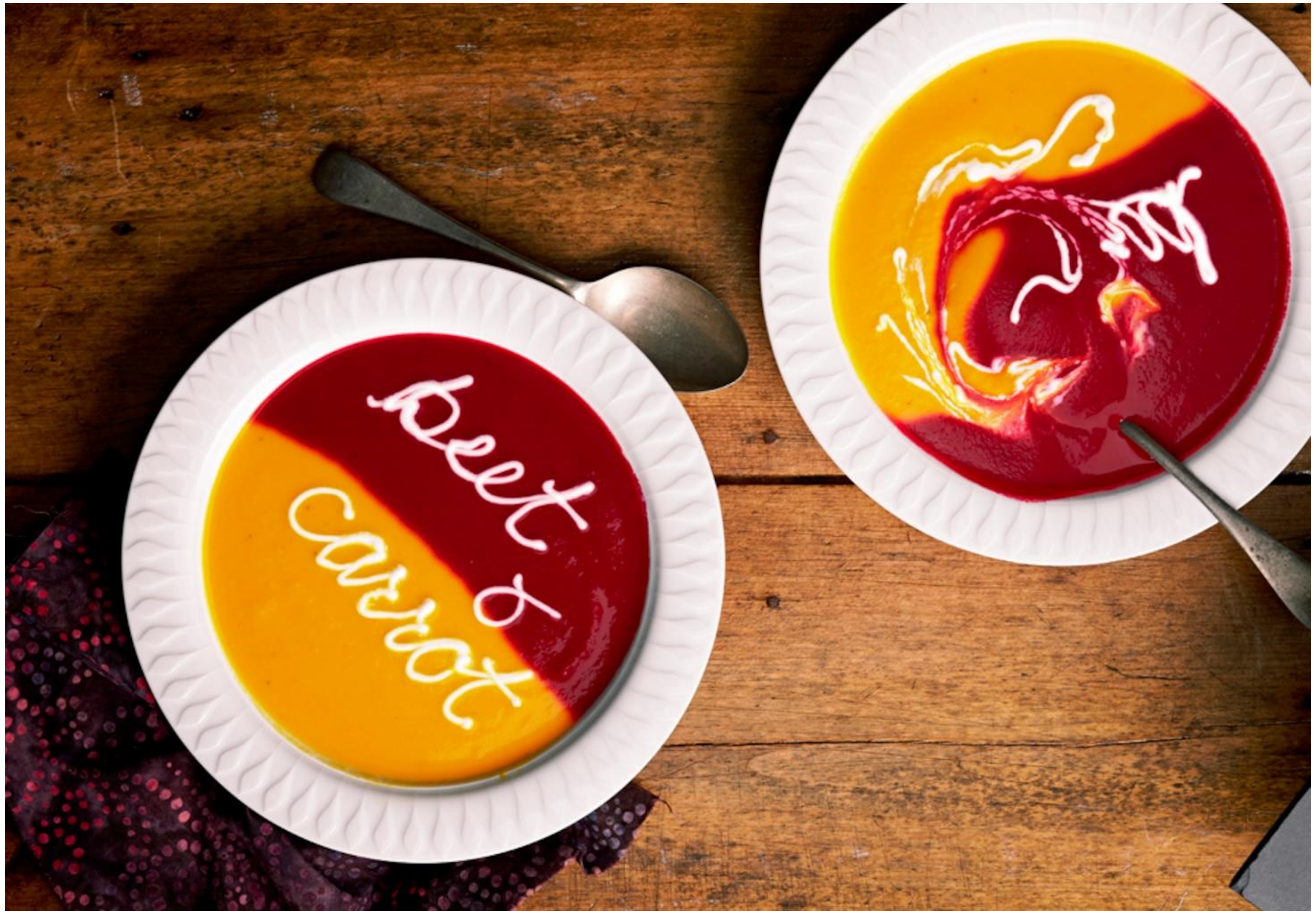
Heat Equation $$\frac{\partial u}{\partial t} = c\Delta u = \mathrm{div}(c\nabla u)$$

SDE formulation (how each particle moves): $$dx(t) = c dW(t)$$

How to do edge preserving filter?

How do we keep the two soups separate?

Why does gaussian filter destroy the edges?



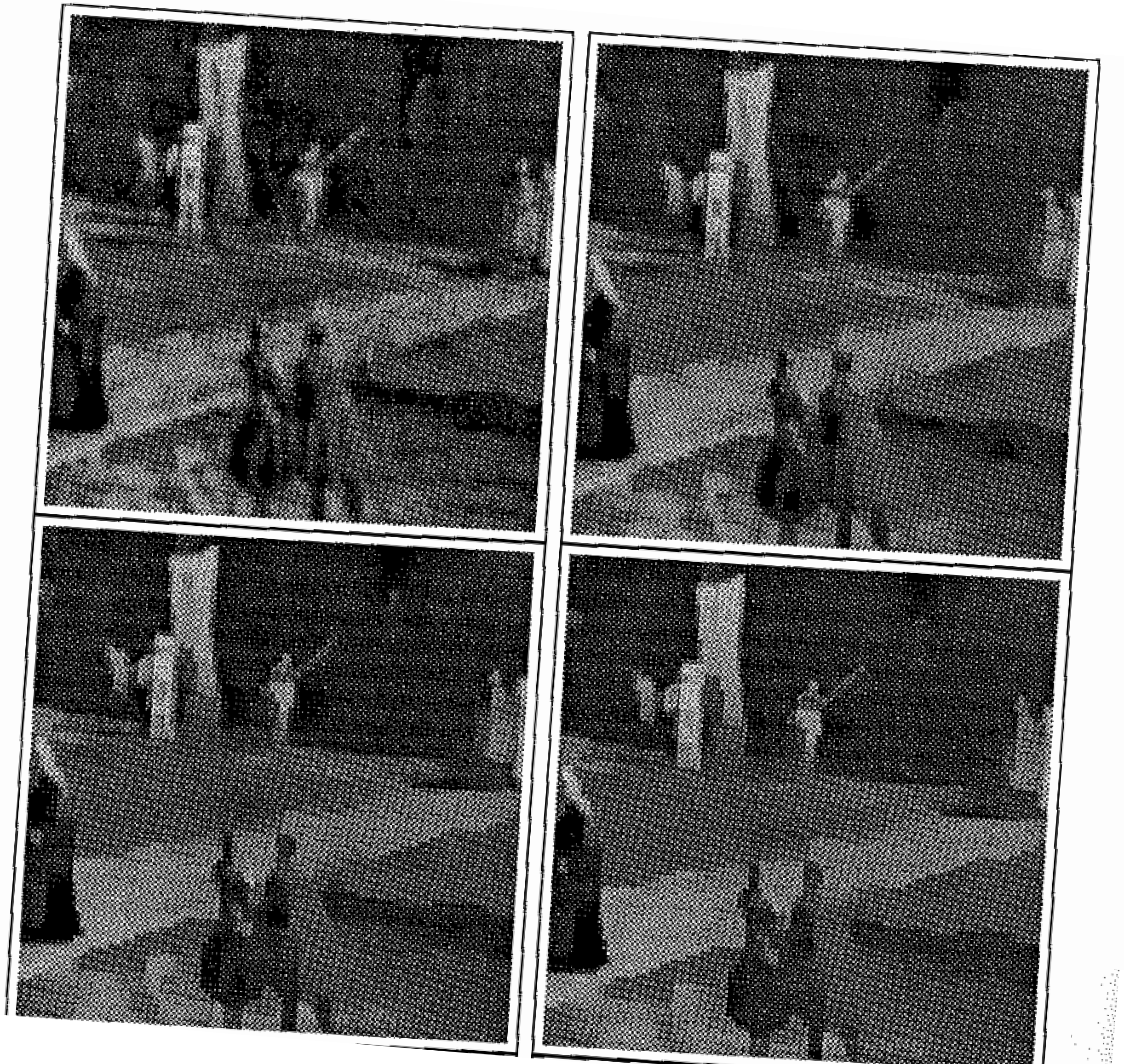
Anisotropic Diffusion Perona & Malik 1990

- Idea: Look to see if there is a wall, modulate diffusion across the edge!
- Anisotropic Diffusion:
  - Treats the edges in the image like this wall

$$\frac{\partial u}{\partial t} = \operatorname{div}(c(|\nabla u|)\nabla u)$$

  - Makes the diffusion process, edge dependent
  - ➔ **Data dependent** filtering

Anisotropic Diffusion Results



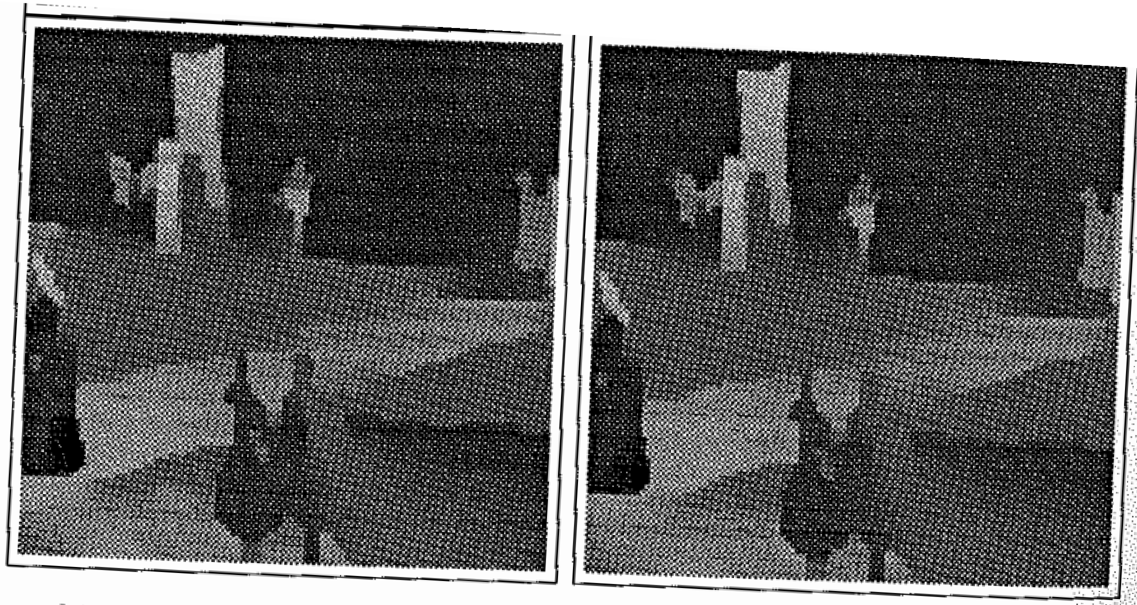


Figure 3.7. Sequence of images produced by anisotropic diffusion. The code presented in figure 3.5 was run on the image at the top-left corner for 10, 20, 30, 60, 100 iterations. The original image has pixel values between 0 (black) and 255 (white) and had a size of 100 × 100 pixels. The coefficient K was set equal to $K = 10$.

Aurich and Weule 1995
Tomasi and Manduchi 1998…

Bilateral Filter

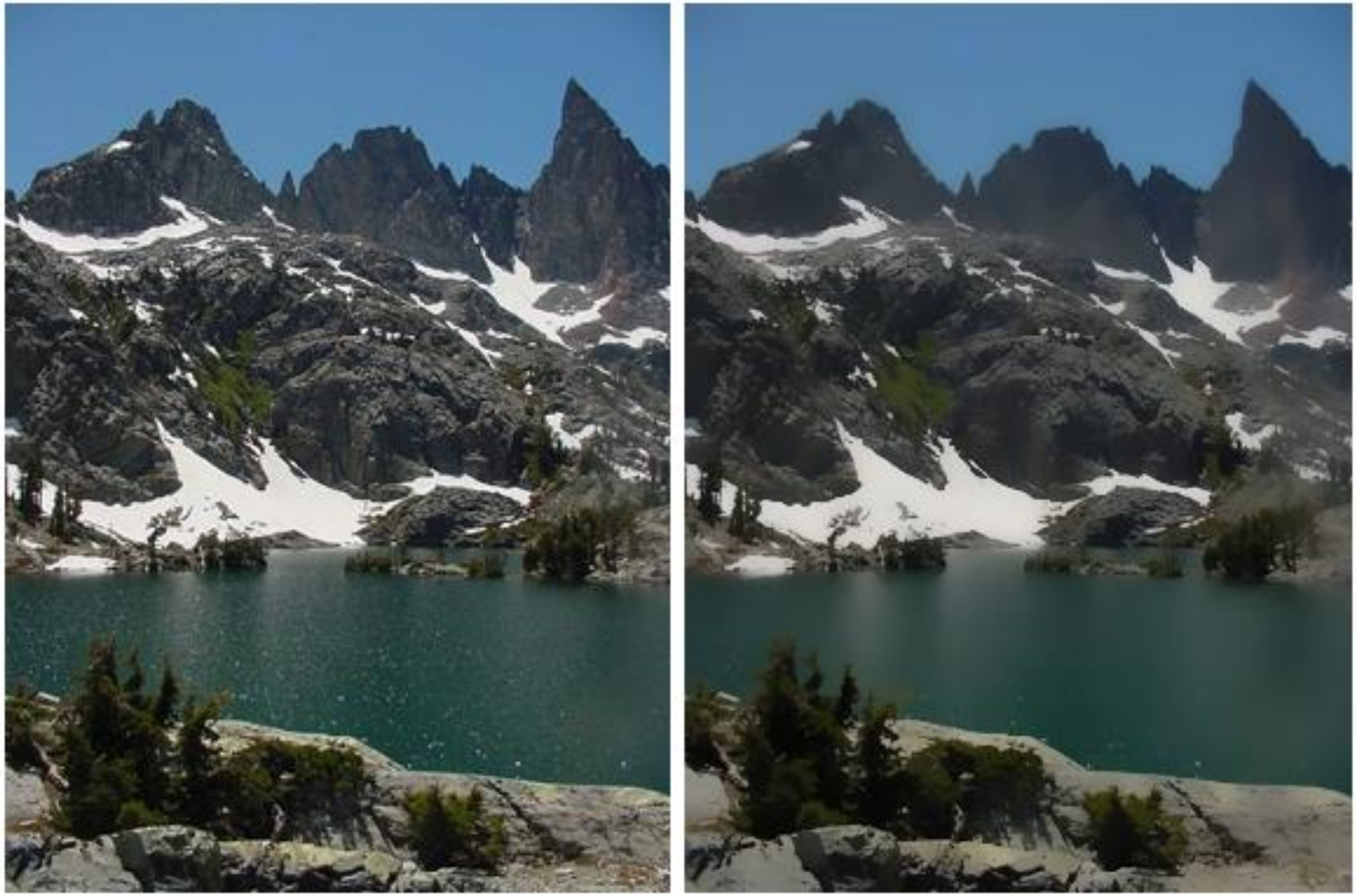
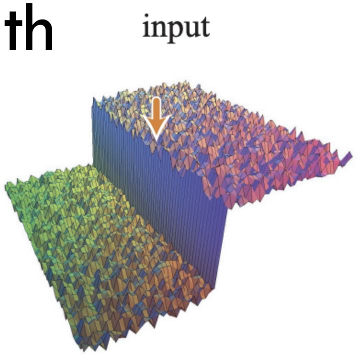


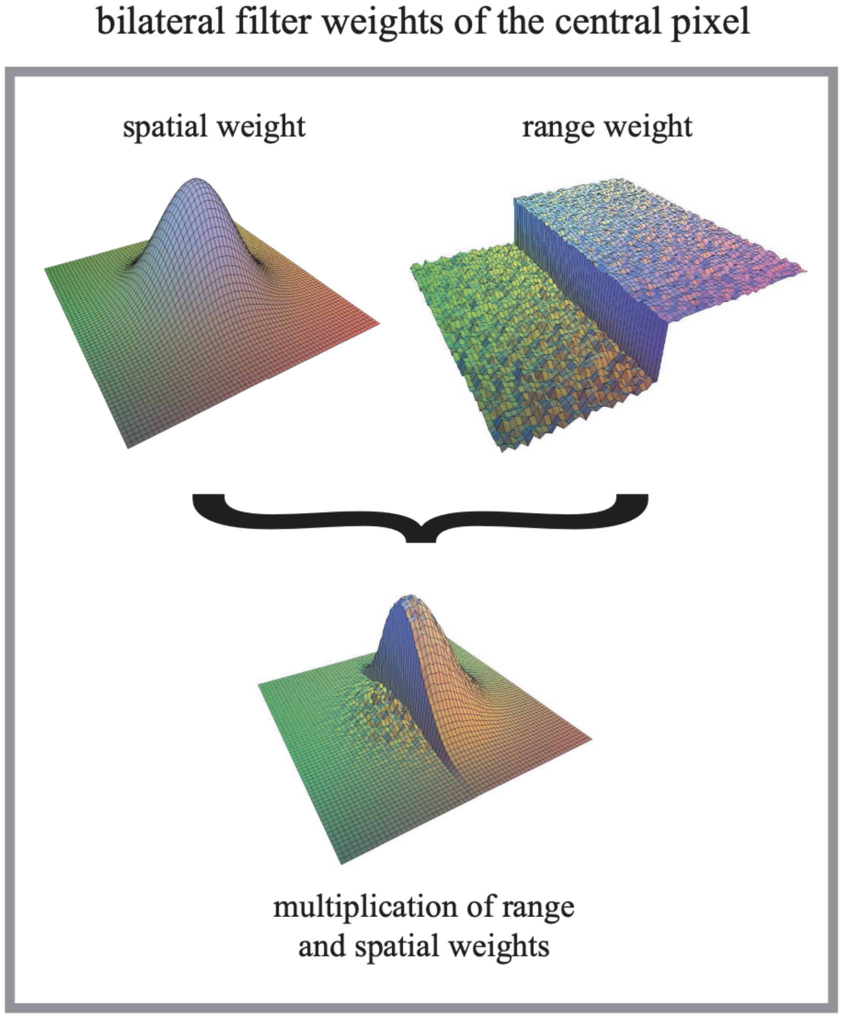
Photo from wikipedia

Aurich and Weule 1995
Tomasi and Manduchi 1998…

Bilateral Filter

- Inspired by Anisotropic Diffusion
- Weight gaussian kernel with pixel similarity
- Again data-dependent
- This is core idea of ROPE




bilateral filter weights of the central pixel
spatial weight
range weight
multiplication of range and spatial weights

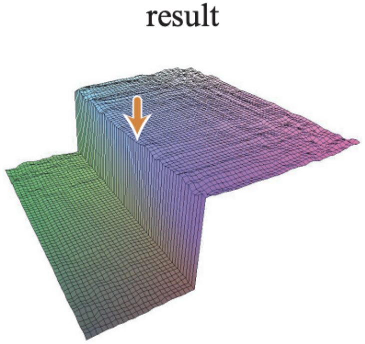
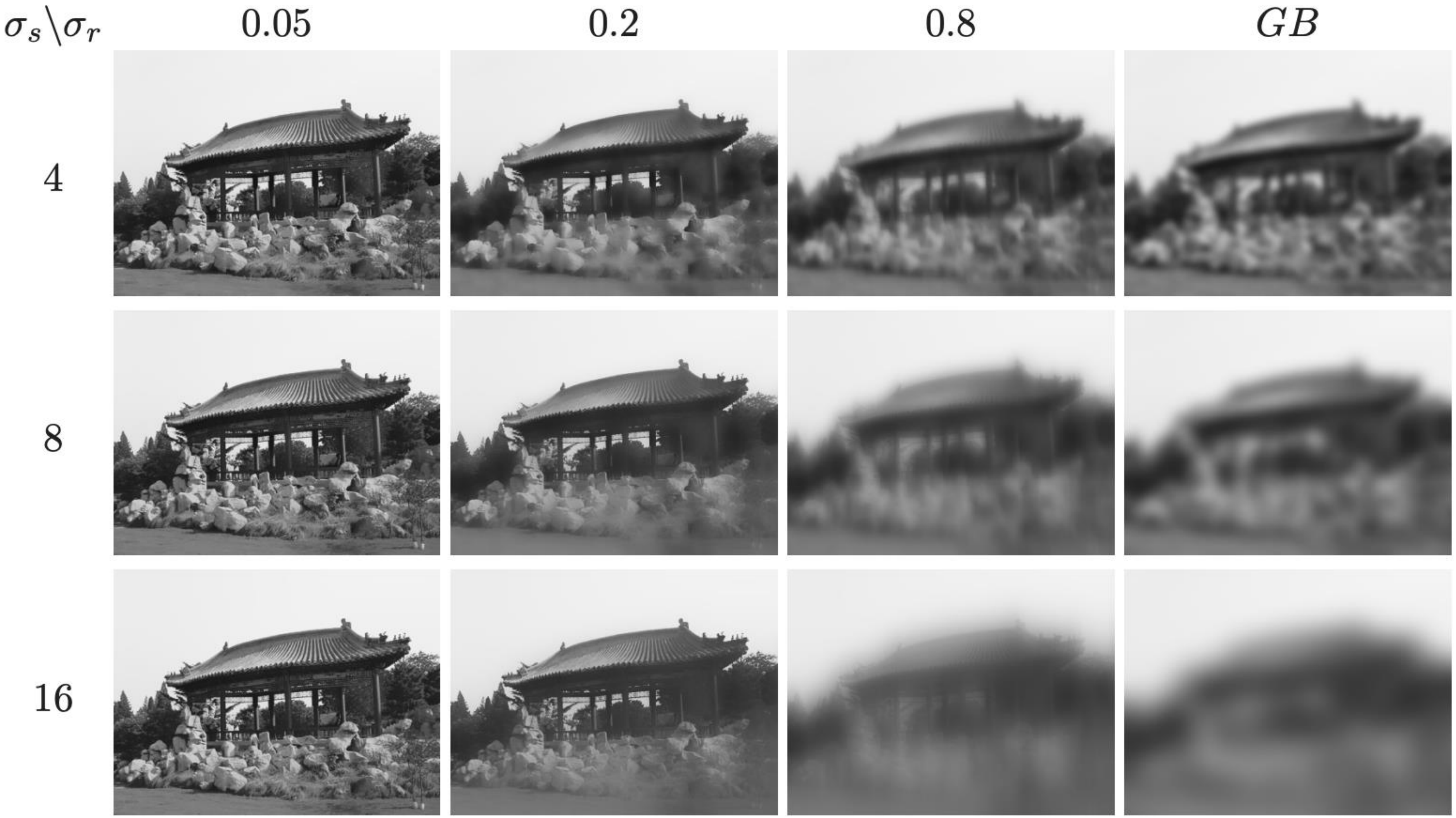


Figure from paris et al.

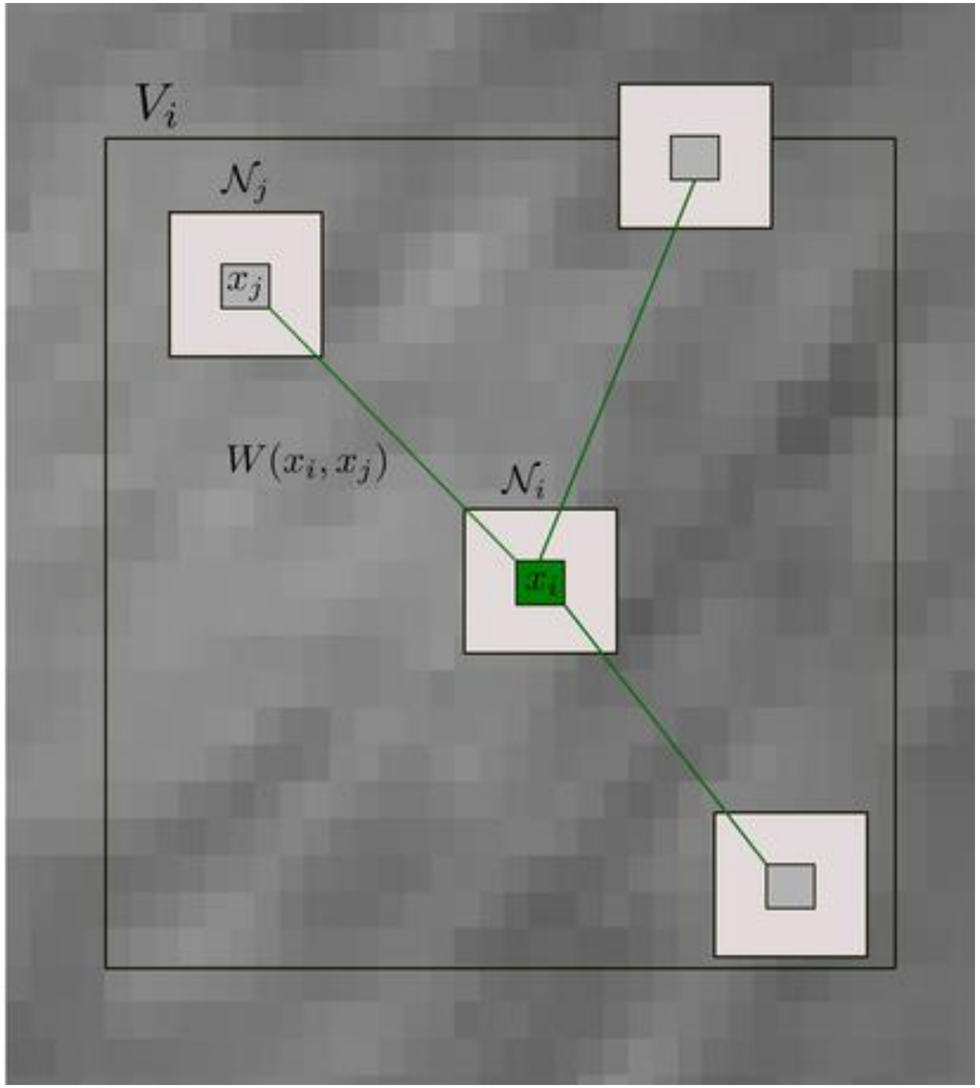
$$BF[I]_{\mathbf{p}} = \frac{1}{W_{\mathbf{p}}} \sum_{\mathbf{q} \in \mathcal{S}} \overbrace{G_{\sigma_s}(\|\mathbf{p}-\mathbf{q}\|)}^{\text{Distance Similarity}} \, \underbrace{G_{\sigma_r}(I_{\mathbf{p}} - I_{\mathbf{q}})}_{\text{Pixel Similarity}} \, I_{\mathbf{q}}$$

Bilateral Filter Results



Non-local means Buades, Coll, and Morel 2005

- Idea of a filter is to denoise by averaging similar pixels
- Why look at near by pixels only? The similar pixels **can be anywhere**!
- Idea: filter by average of similar patches, **from everywhere in the image**!
- Data-driven weights (similarity) with large field of view.



Non-local means

$$NL[v](i) = \sum_{j \in I} w(i,j) v(j),$$

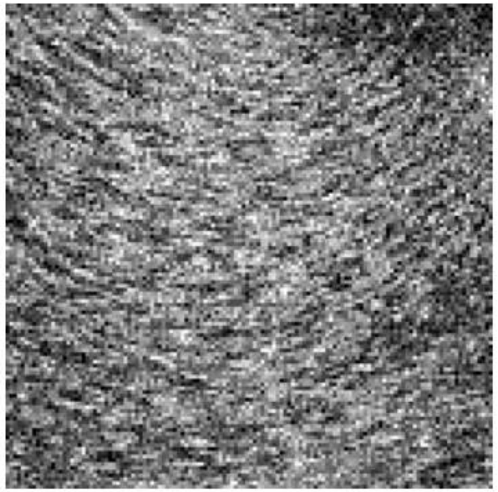
Weight is normalized affinity to all other pixels:

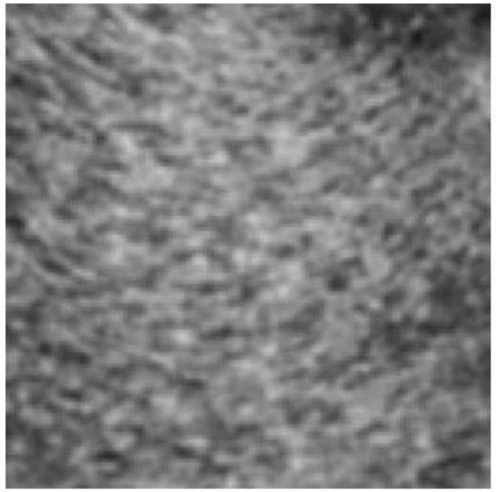
$$w(i,j) = \frac{1}{z_i} e^{-\|v_i - v_j\|/\sigma^2} \qquad z_i = \sum_j e^{-\|v_i - v_j\|/\sigma^2}$$

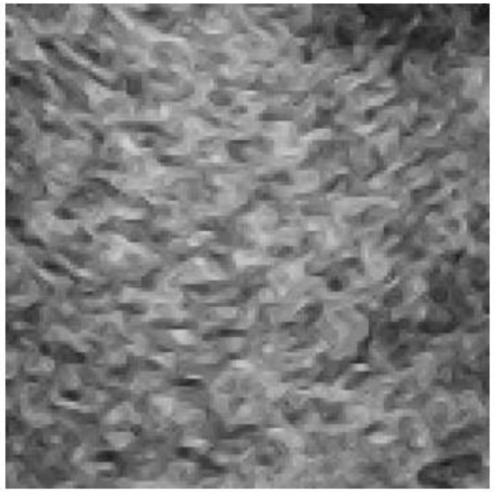
Softmax! In vectorized form, this is:

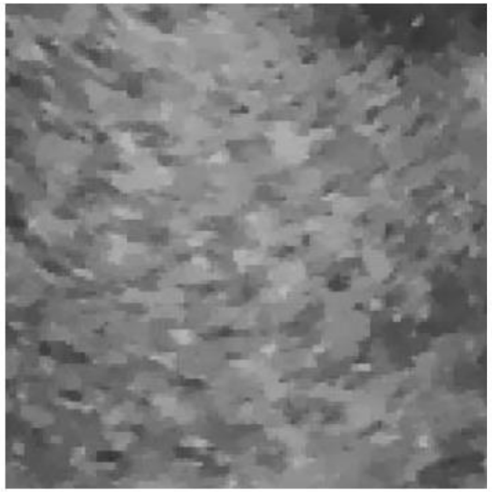
$$NL = softmax\big(dist(v_i, v_j)\big)V$$

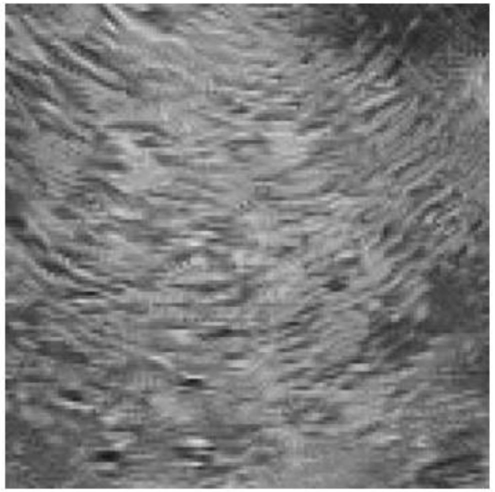
Results


Input


Gaussian Filter

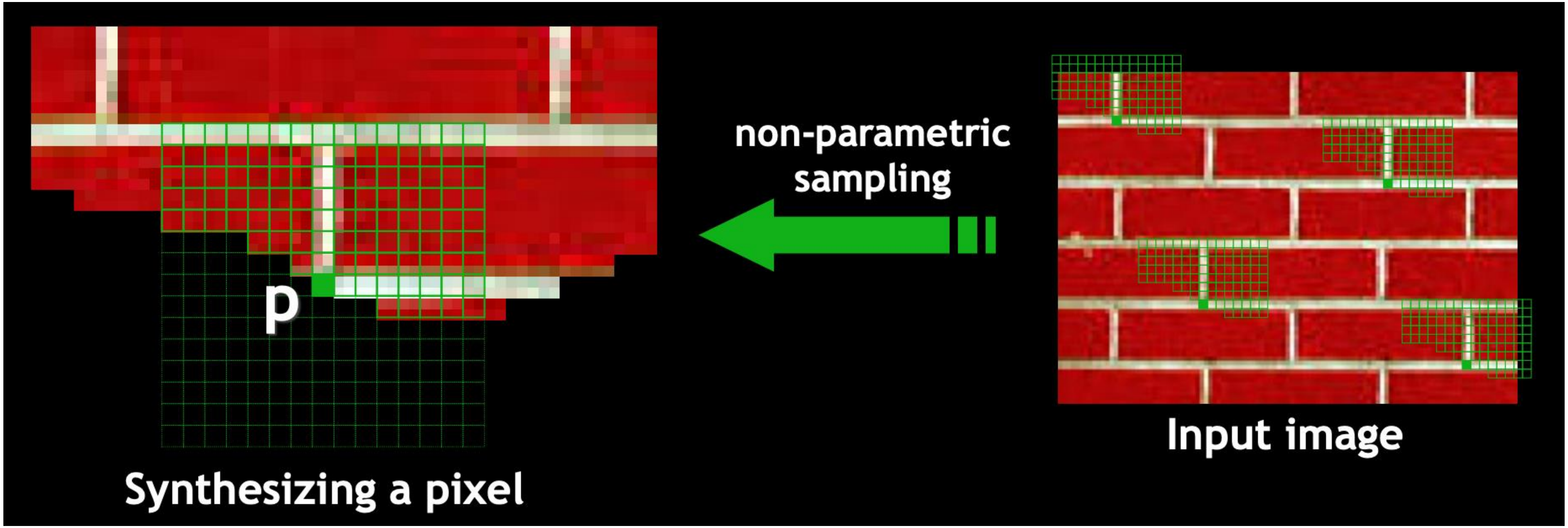

Anisotropic Filter


Total Variation


Non-local means

Aside Efros and Leung 1999

- Inspired non-local means



Two key ideas

In contrast to ordinary convlution by a fixed kernel

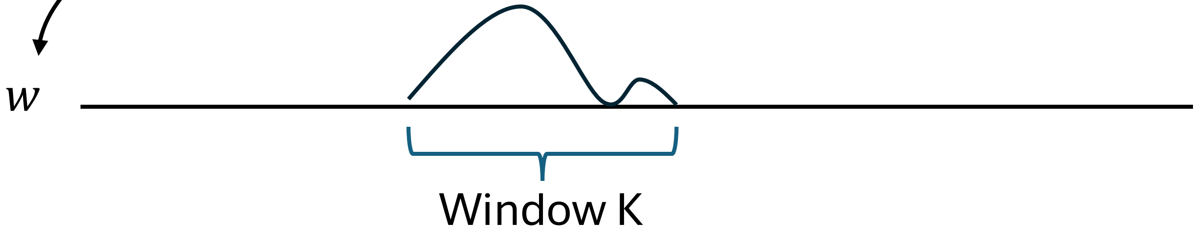
1. **Data-driven kernel** → Weights conditioned on the data point
2. **Full field of view** → non-local connections

These are the key difference between convnets and transformers

Convolution:

$$\boldsymbol{x}' = W\boldsymbol{x} + \boldsymbol{b}$$

$$x_i' = \sum_{j}^{N} w_j x_j$$



Attention:

$$\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

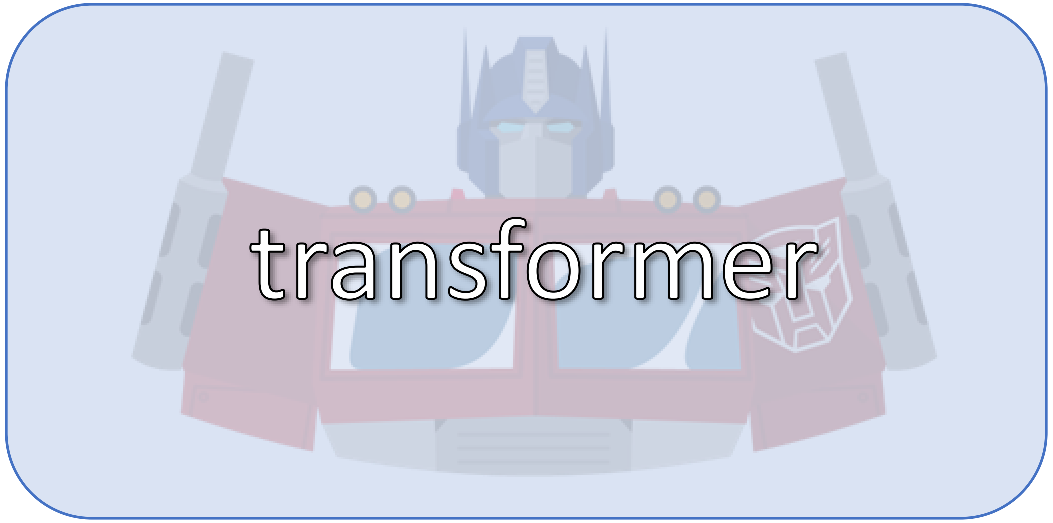
$$x_i' = \sum_{j}^{N} w_j(x_i, x_j)x_j$$

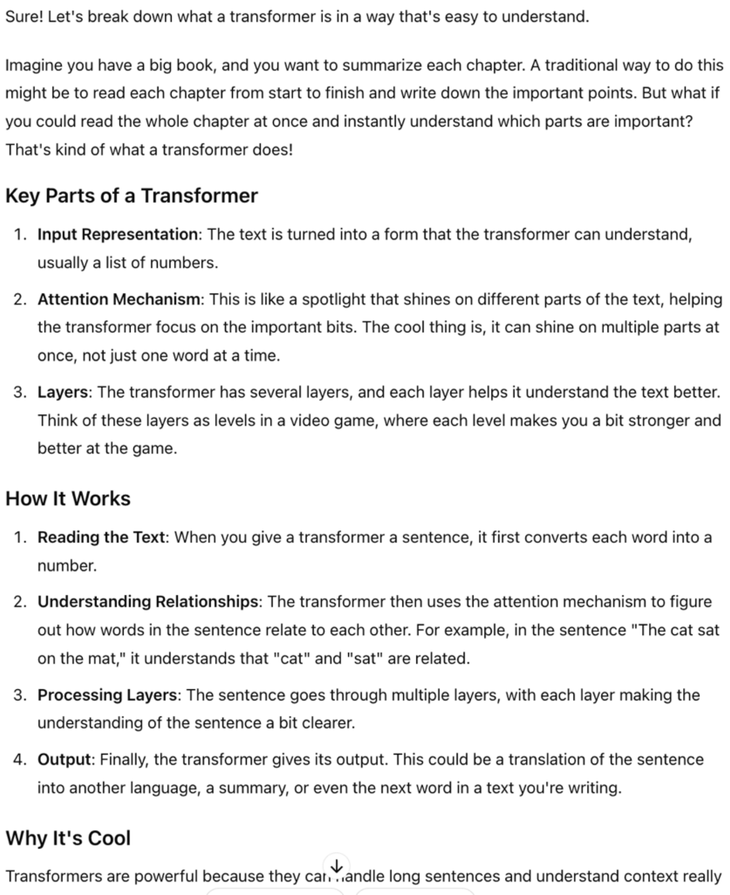
Transformer Architecture

Excellent slides from Jiabin Huang (UMD) for the rest

https://www.youtube.com/watch?v=rcWMRA9E5RI

Please explain transformers like I am five





Sure! Let's break down what a transformer is in a way that's easy to understand.

Imagine you have a big book, and you want to summarize each chapter. A traditional way to do this might be to read each chapter from start to finish and write down the important points. But what if you could read the whole chapter at once and instantly understand which parts are important? That's kind of what a transformer does!

Key Parts of a Transformer

1. **Input Representation**: The text is turned into a form that the transformer can understand, usually a list of numbers.
2. **Attention Mechanism**: This is like a spotlight that shines on different parts of the text, helping the transformer focus on the important bits. The cool thing is, it can shine on multiple parts at once, not just one word at a time.
3. **Layers**: The transformer has several layers, and each layer helps it understand the text better. Think of these layers as levels in a video game, where each level makes you a bit stronger and better at the game.

How It Works

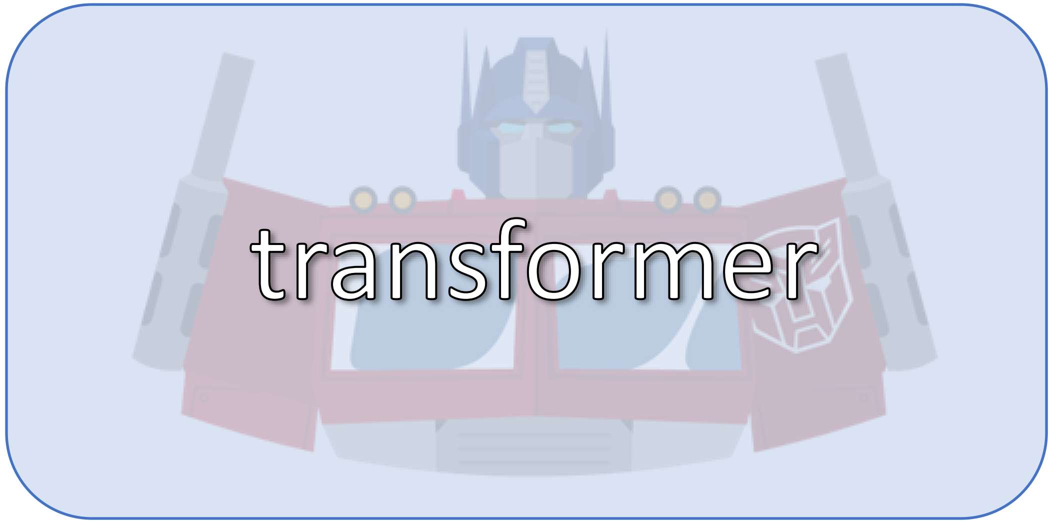
1. **Reading the Text**: When you give a transformer a sentence, it first converts each word into a number.
2. **Understanding Relationships**: The transformer then uses the attention mechanism to figure out how words in the sentence relate to each other. For example, in the sentence "The cat sat on the mat," it understands that "cat" and "sat" are related.
3. **Processing Layers**: The sentence goes through multiple layers, with each layer making the understanding of the sentence a bit clearer.
4. **Output**: Finally, the transformer gives its output. This could be a translation of the sentence into another language, a summary, or even the next word in a text you're writing.

Why It's Cool

Transformers are powerful because they can handle long sentences and understand context really

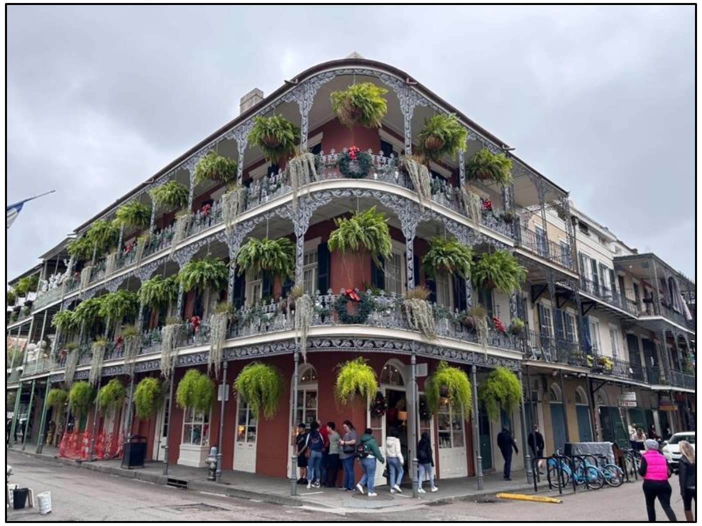
ChatGPT (OpenAI)

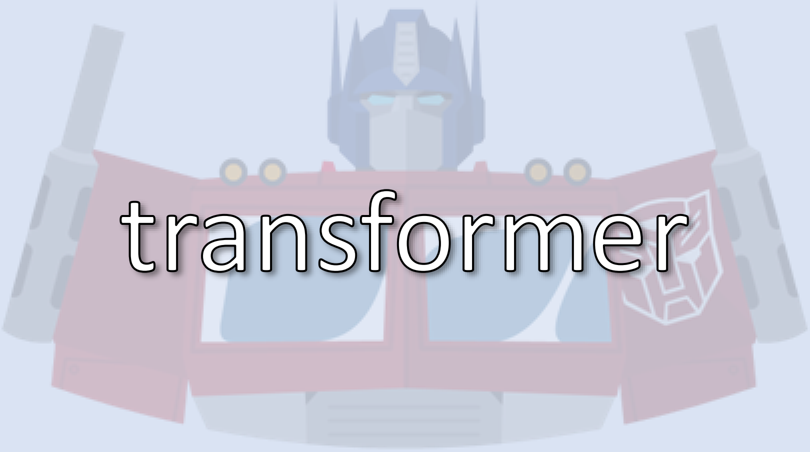
Beautiful, snowy Tokyo city is bustling. The camera moves through the bustling city street, following several people enjoying the beautiful snowy weather and shopping at nearby stalls. Gorgeous sakura petals are flying through the wind along with snowflakes.



Sora (OpenAI)

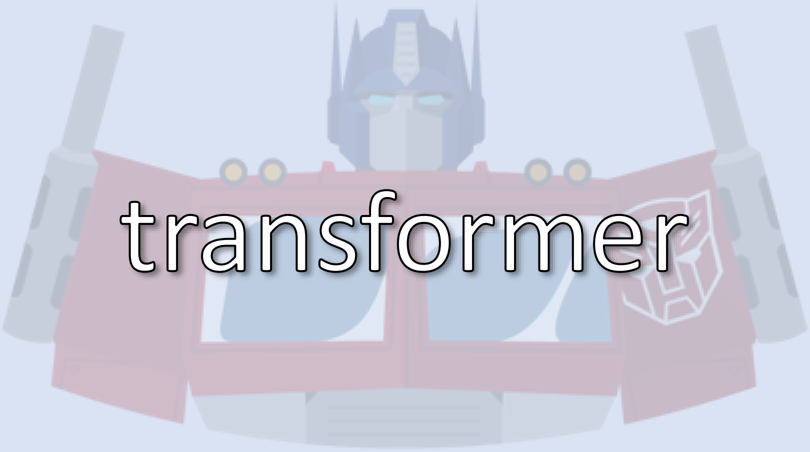
Which street was this photo taken on?





The photo was taken on Bourbon Street, which is located in the French Quarter of New Orleans, Louisiana.

Unified-IO (AI2)

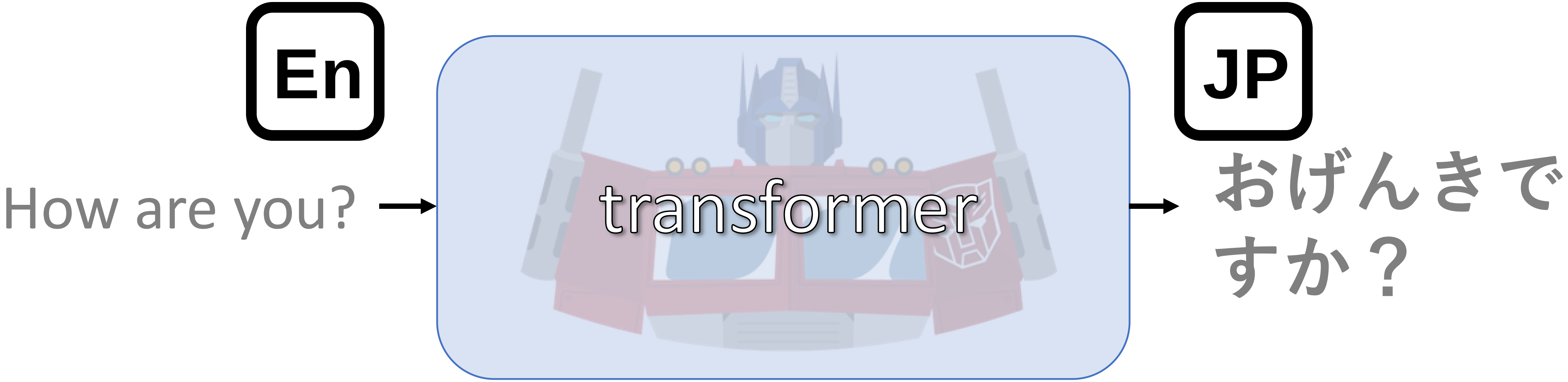


transformer

What?

How?

Why?



En

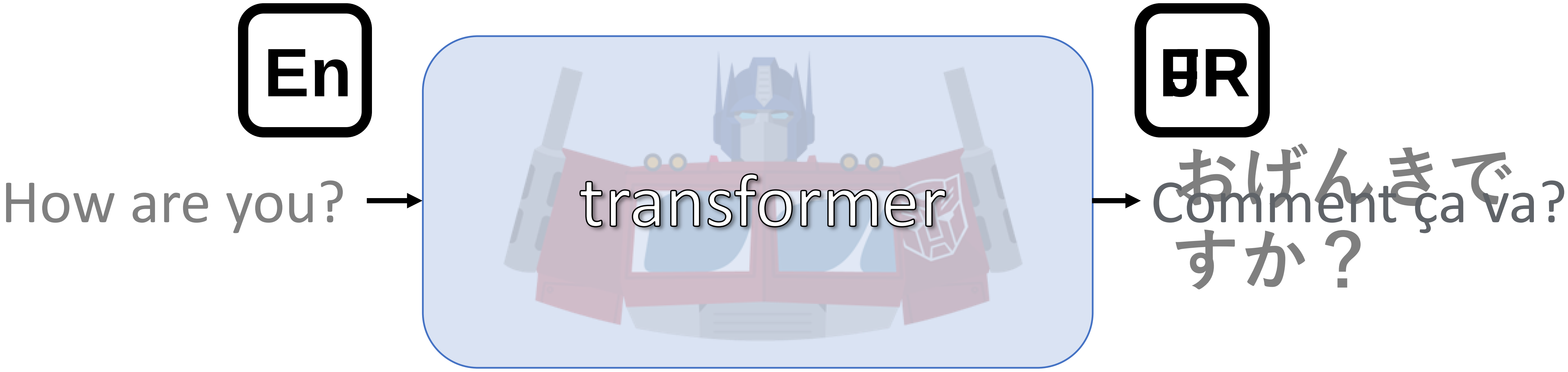
How are you?

transformer

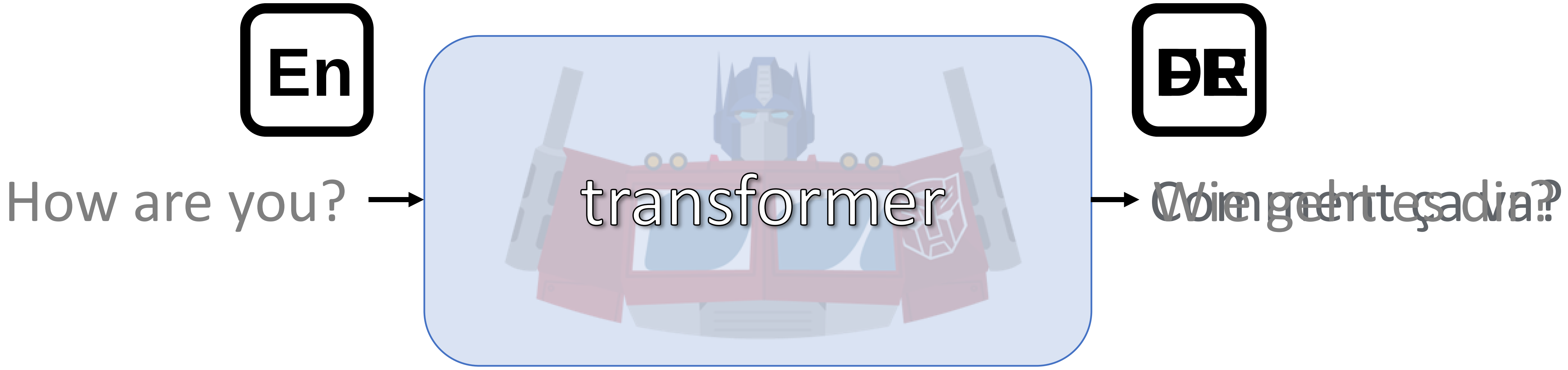
JP

おげんきですか？

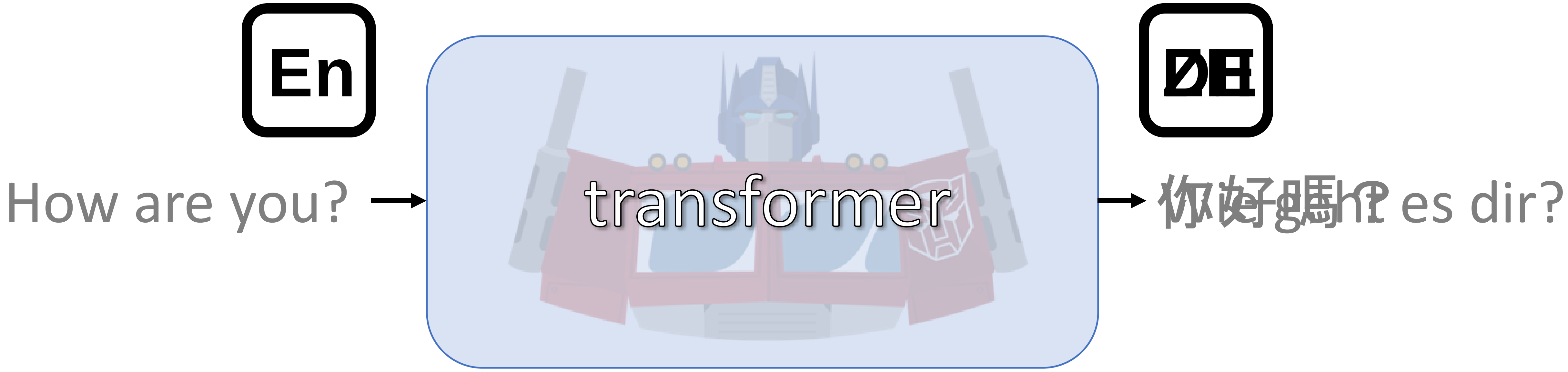
Sequence-to-Sequence model



Sequence-to-Sequence model



Sequence-to-Sequence model



En

How are you?

transformer

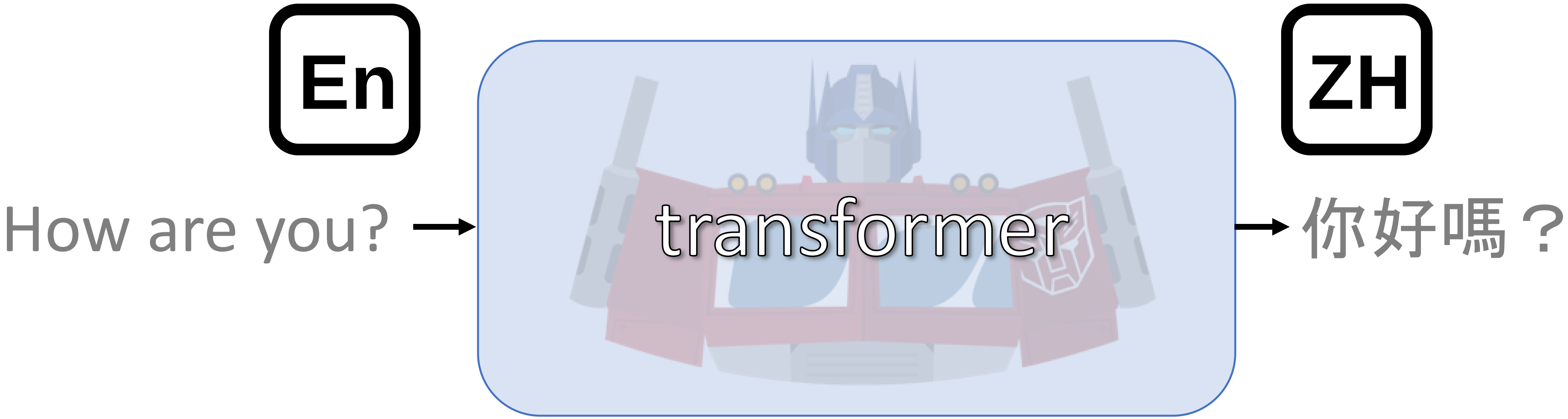
ZH

DE

你好嗎？

Wie geht es dir?

Sequence-to-Sequence model



En

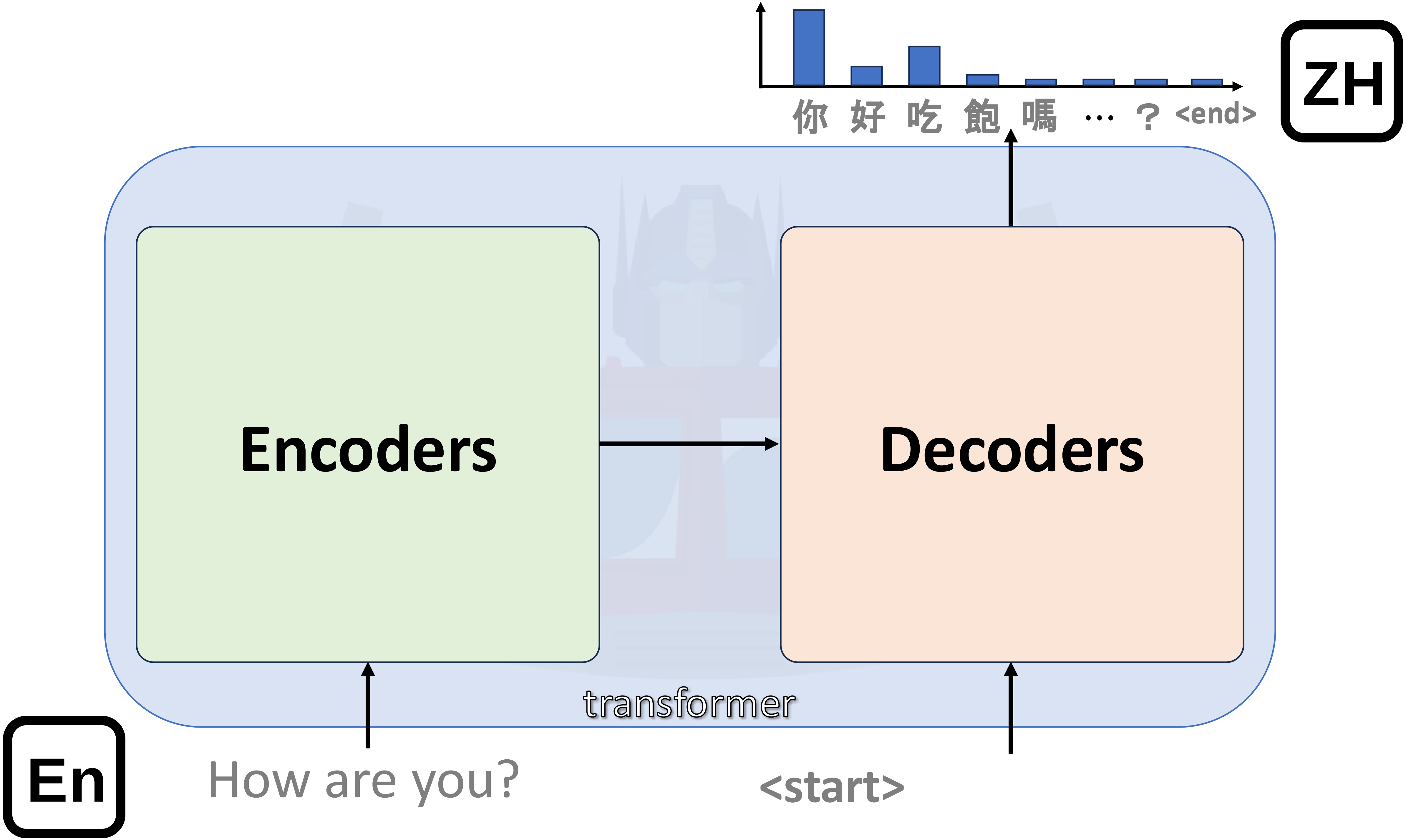
How are you?

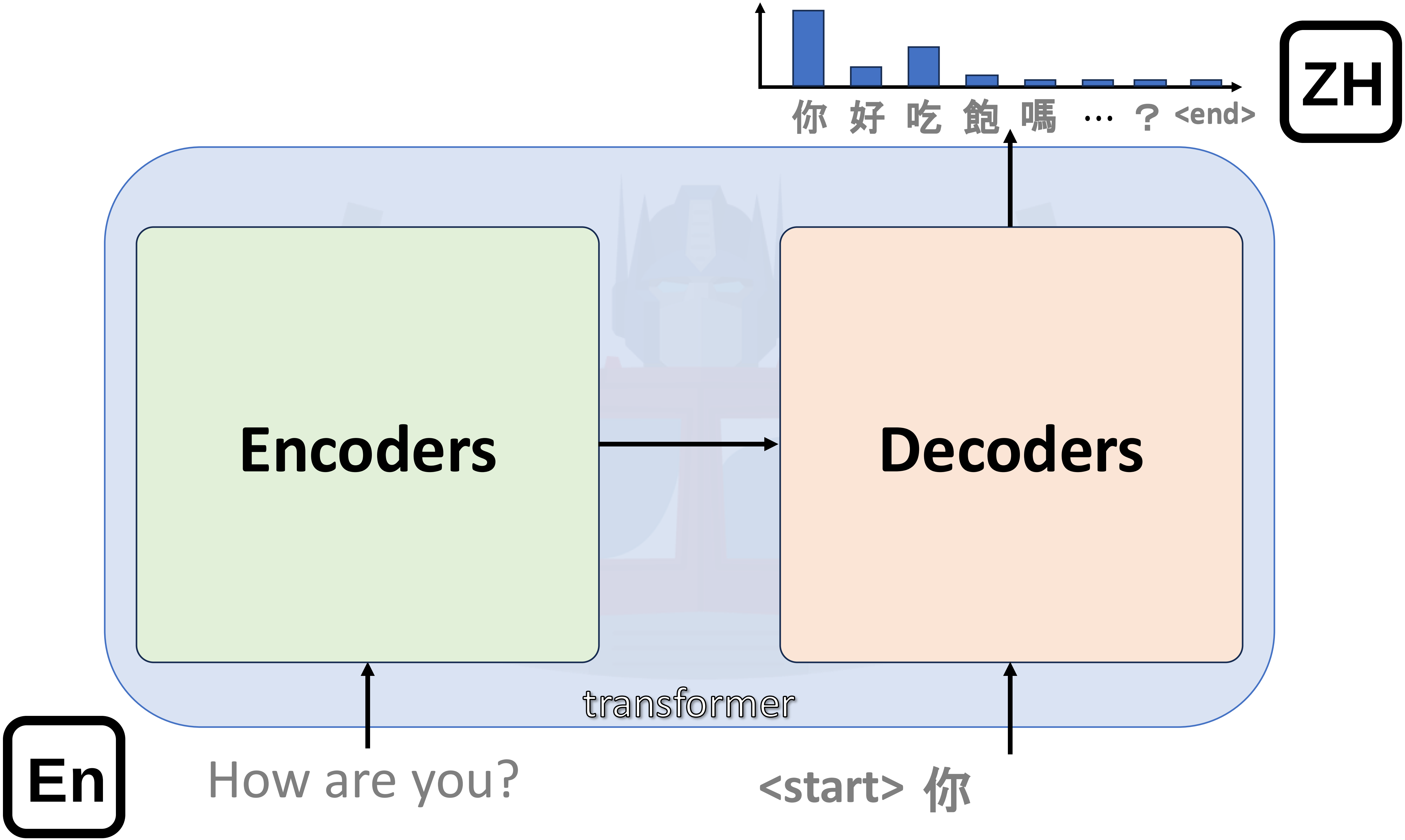
transformer

ZH

你好嗎？

Sequence-to-Sequence model


你
好
吃
飽
嗎
…
？
<end>
ZH
Encoders
Decoders
transformer
En
How are you?
<start>



ZH

Encoders

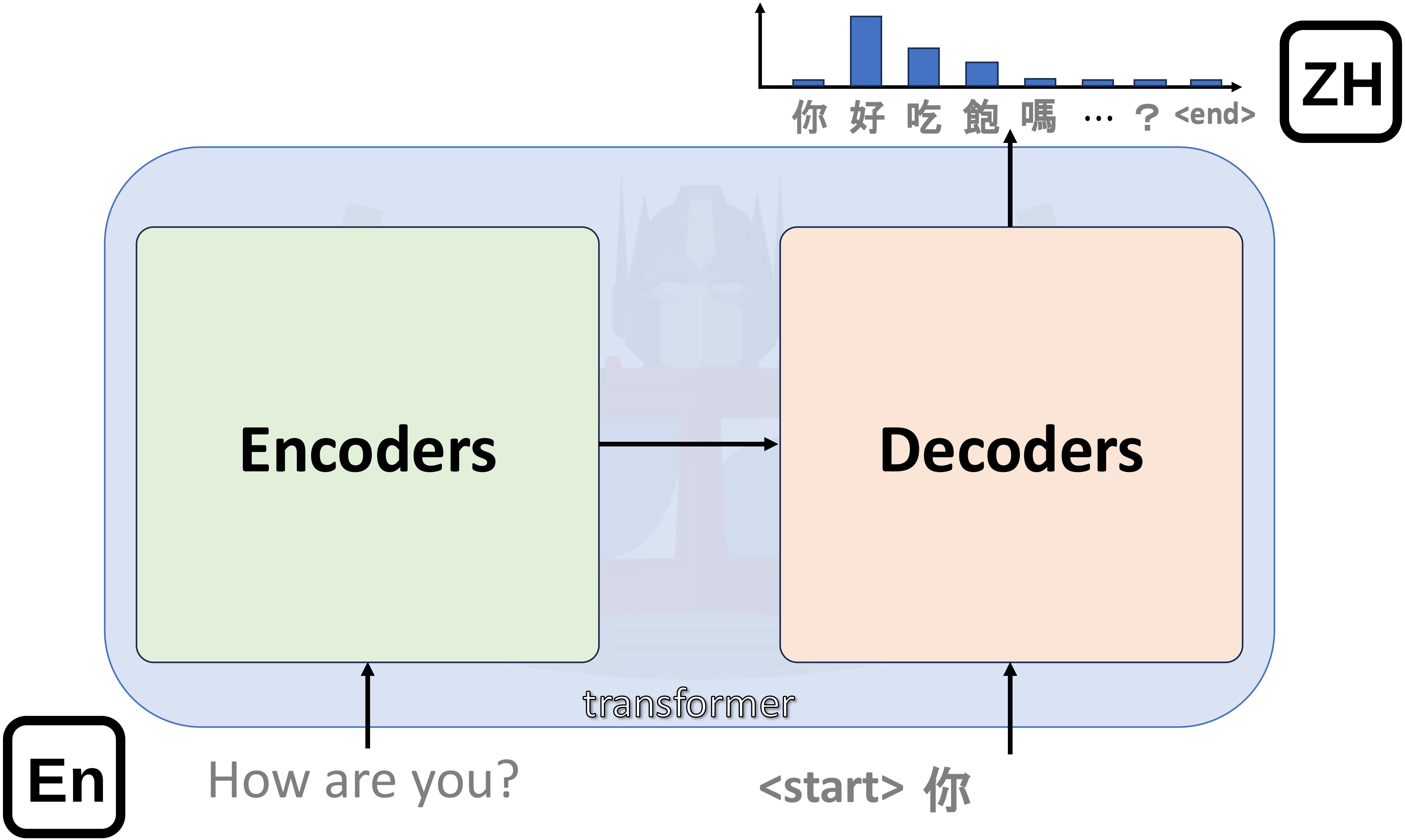
Decoders

transformer

En

How are you?

<start> 你



ZH

Encoders

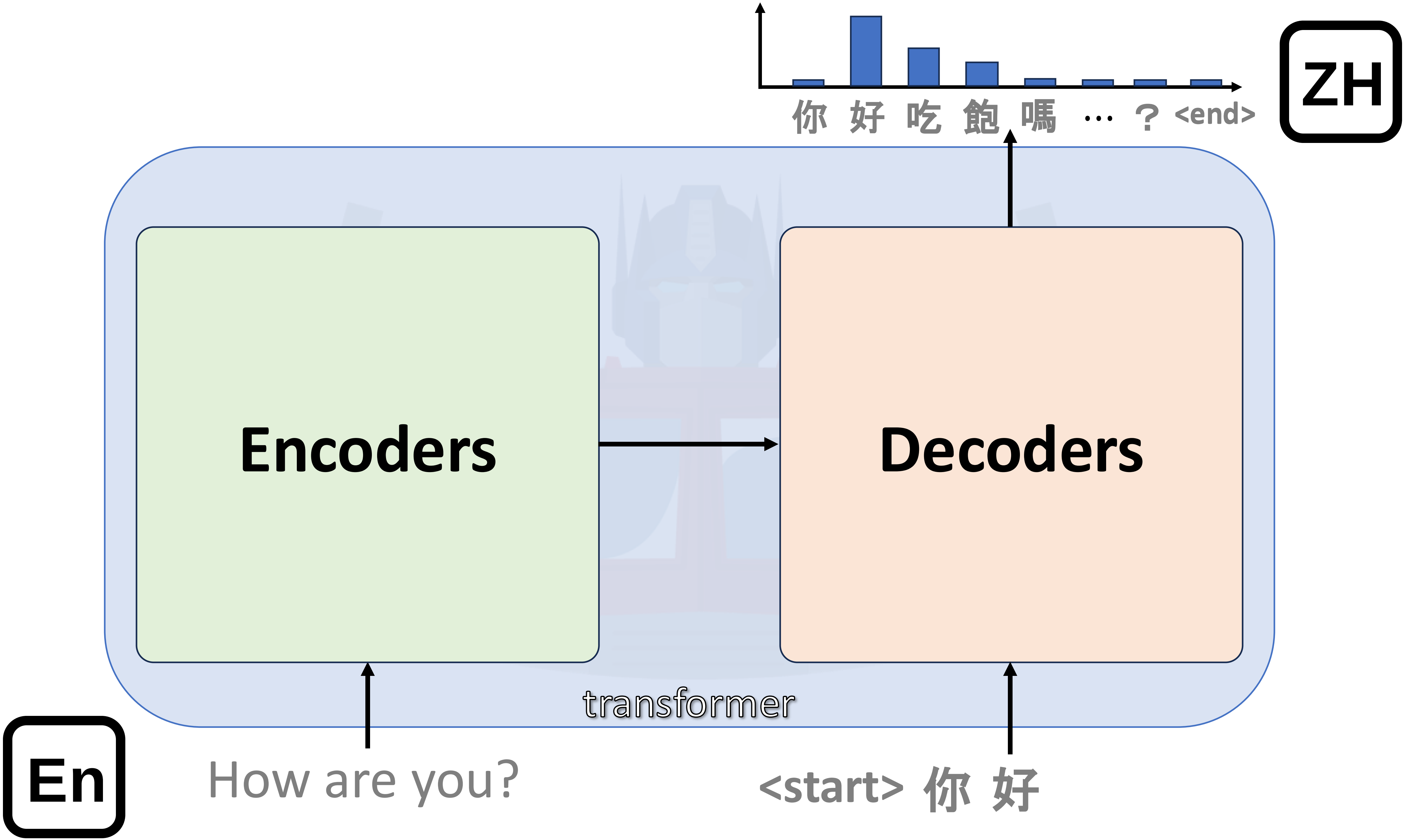
Decoders

transformer

En

How are you?

<start> 你



ZH

Encoders

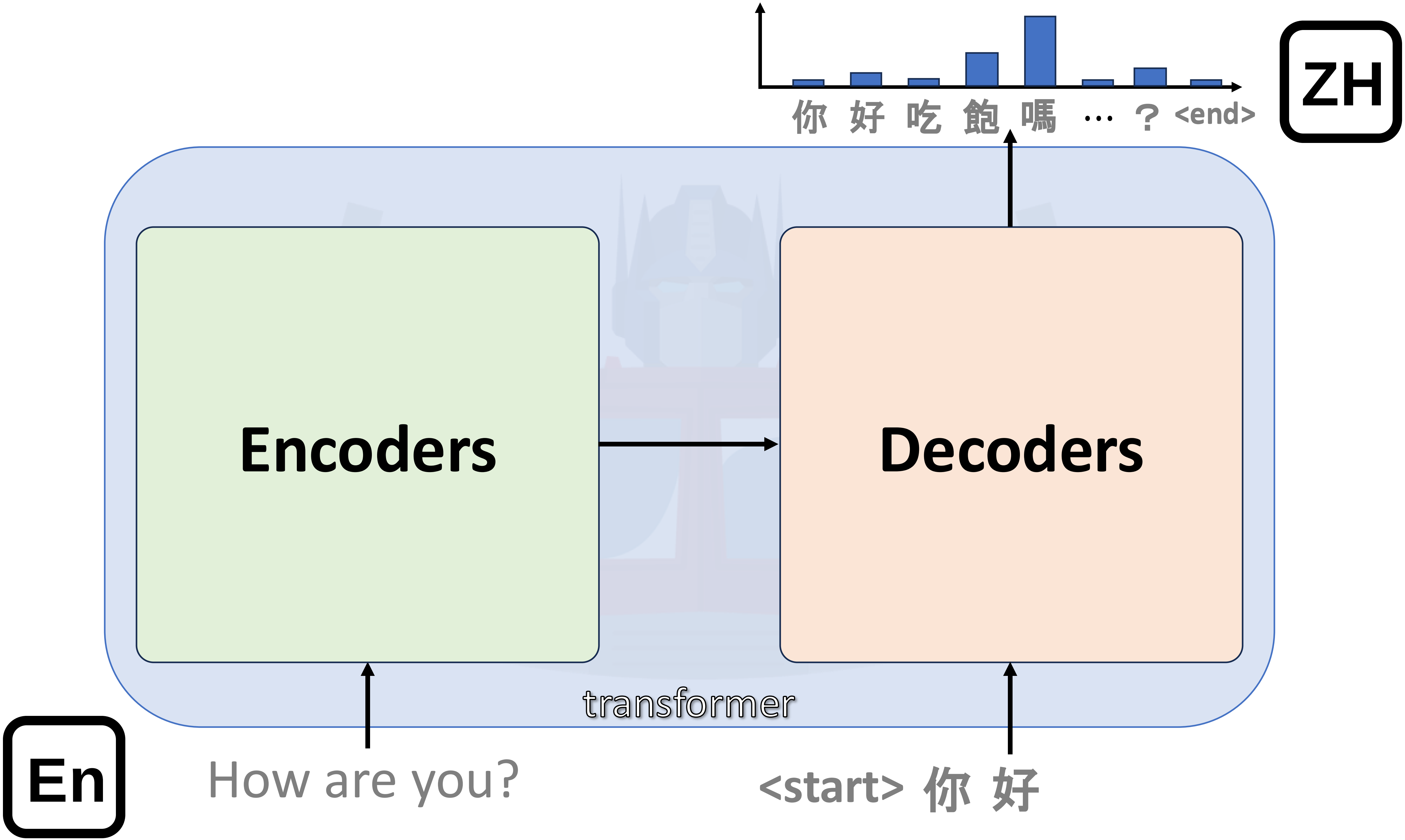
Decoders

transformer

En

How are you?

<start> 你 好



ZH

Encoders

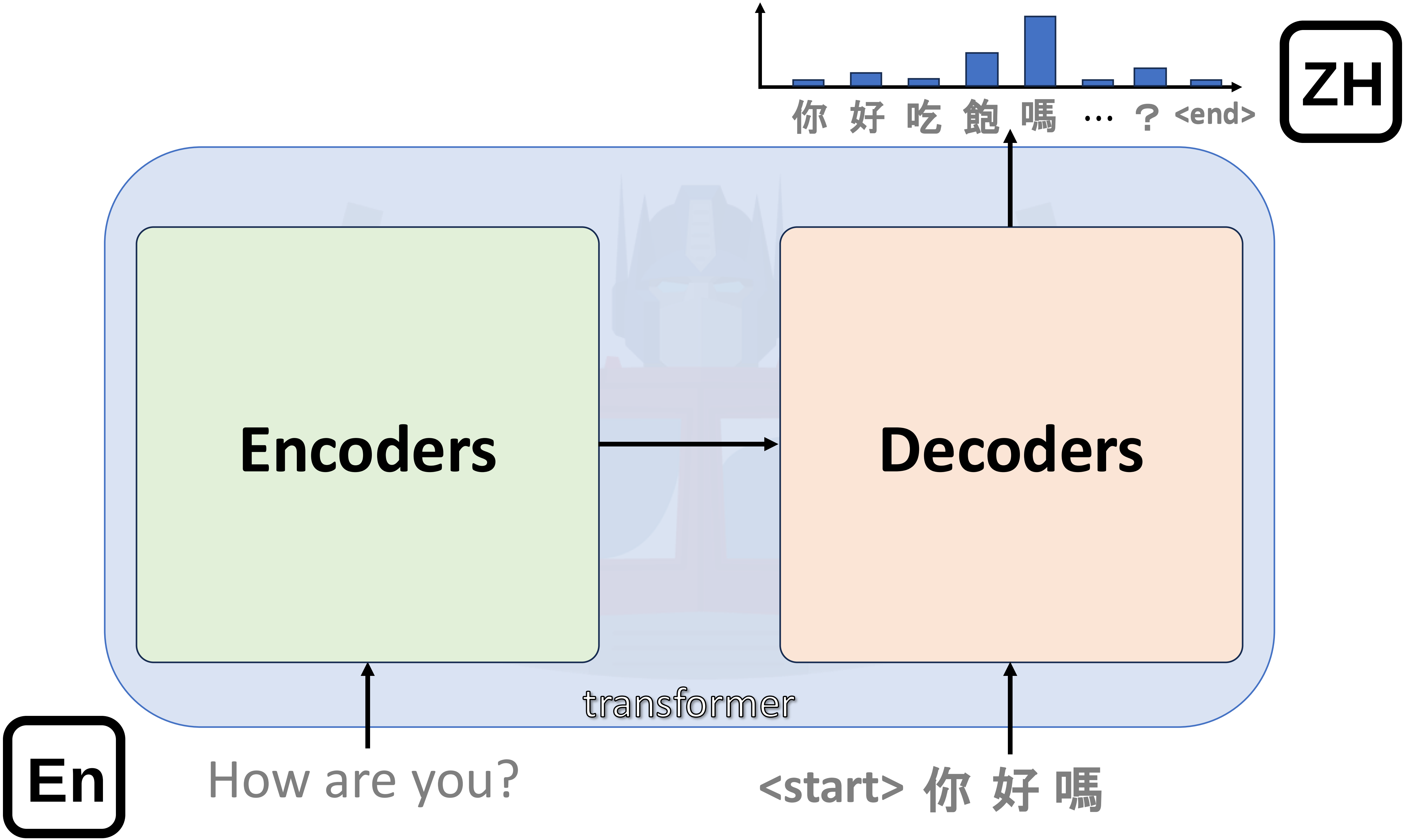
Decoders

transformer

En

How are you?

<start> 你 好



Encoders

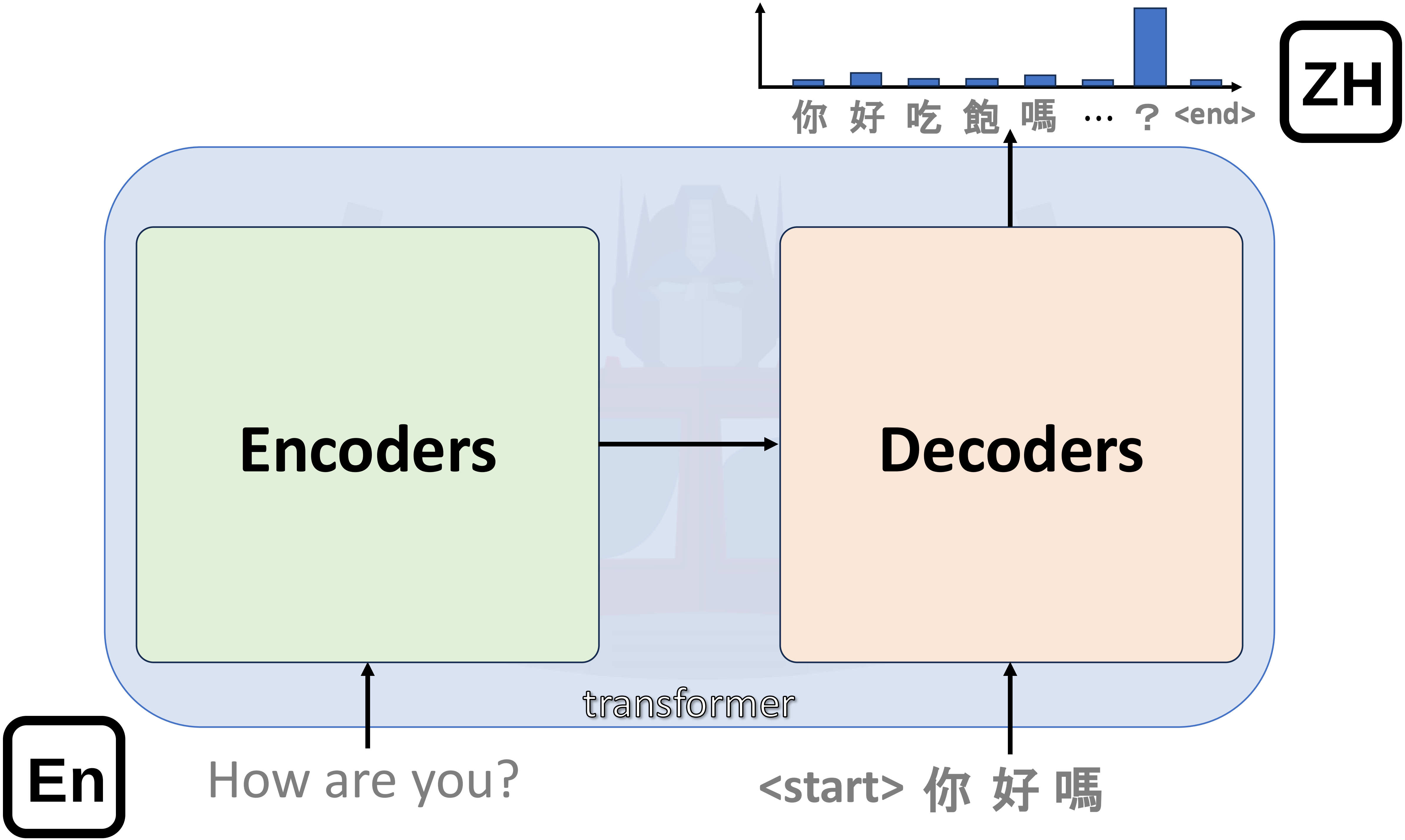
Decoders

transformer

En

How are you?

<start> 你 好 嗎



ZH

Encoders

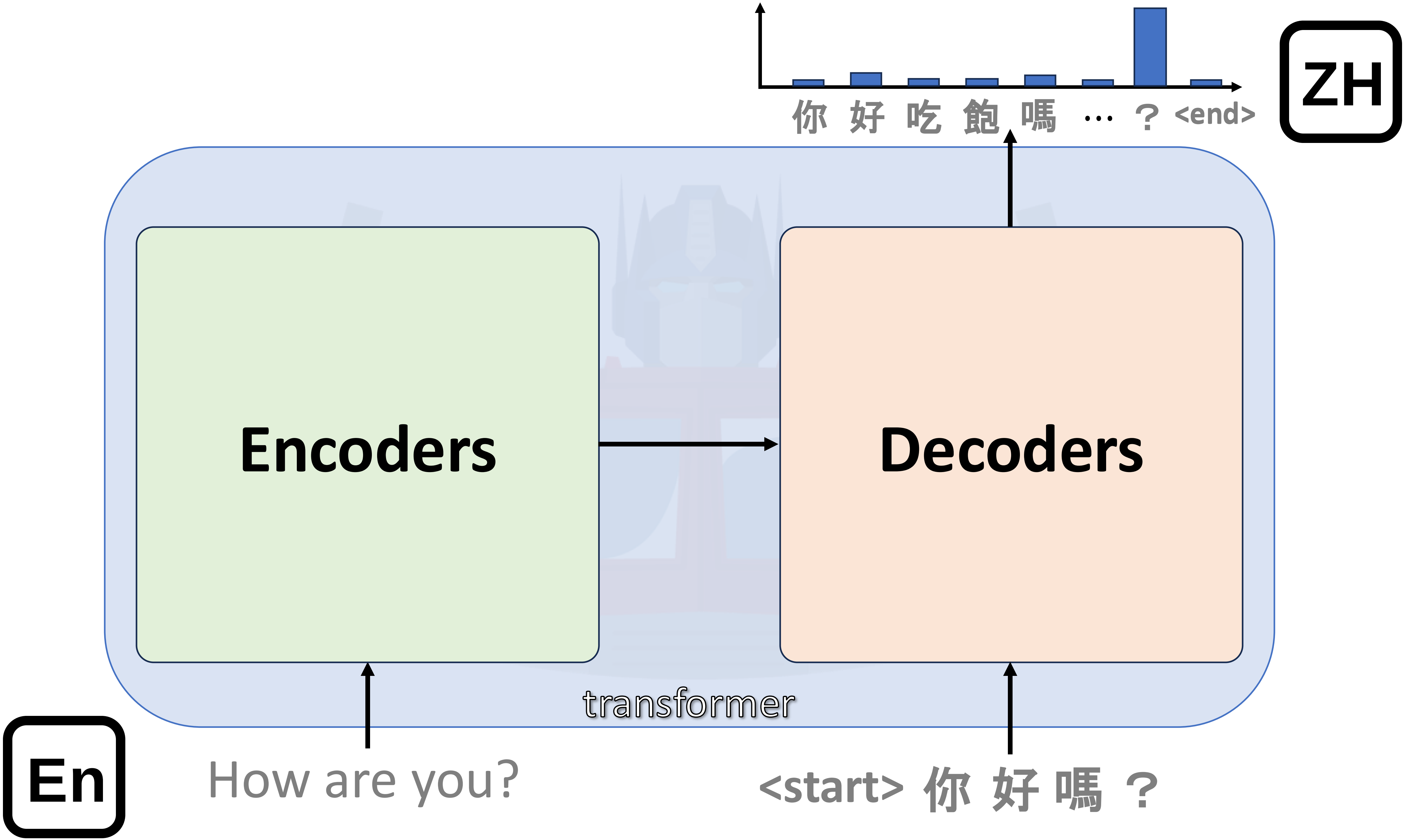
Decoders

transformer

En

How are you?

<start> 你 好 嗎



ZH

你 好 吃 飽 嗎 … ？ <end>

Encoders

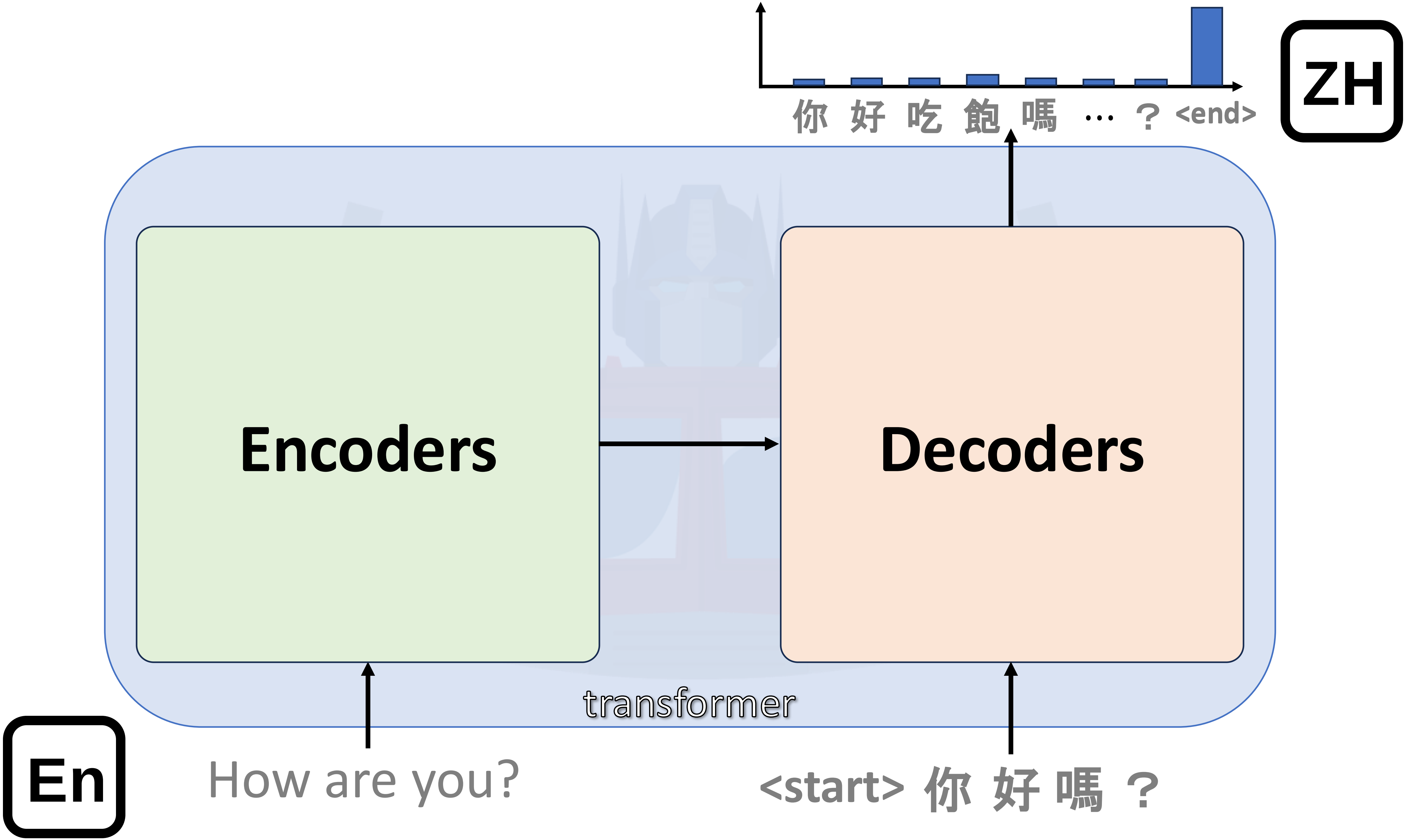
Decoders

transformer

En

How are you?

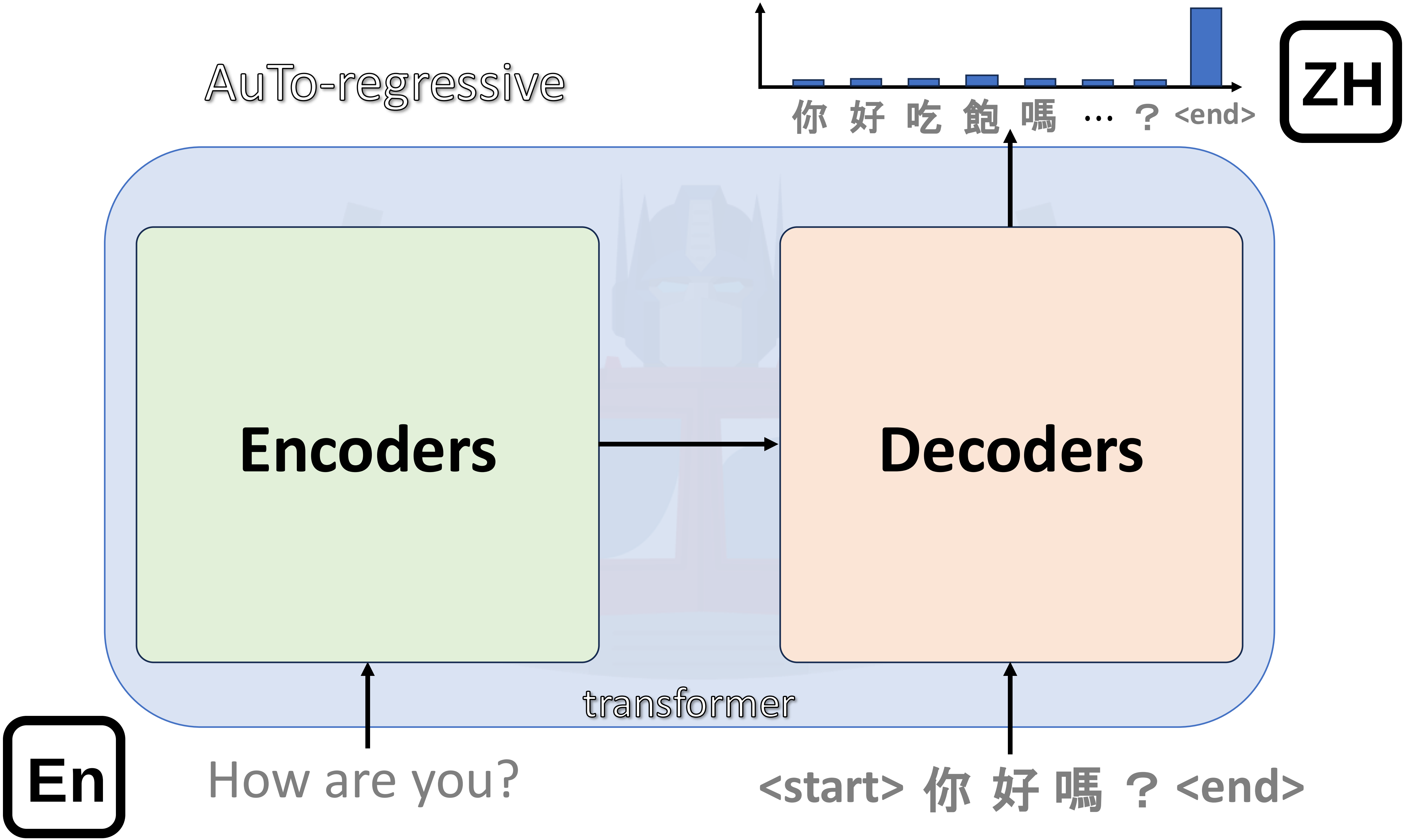
<start> 你 好 嗎 ？



How are you?

<start> 你 好 嗎 ？

AuTo-regressive



你 好 吃 飽 嗎 … ？ <end>

ZH

Encoders

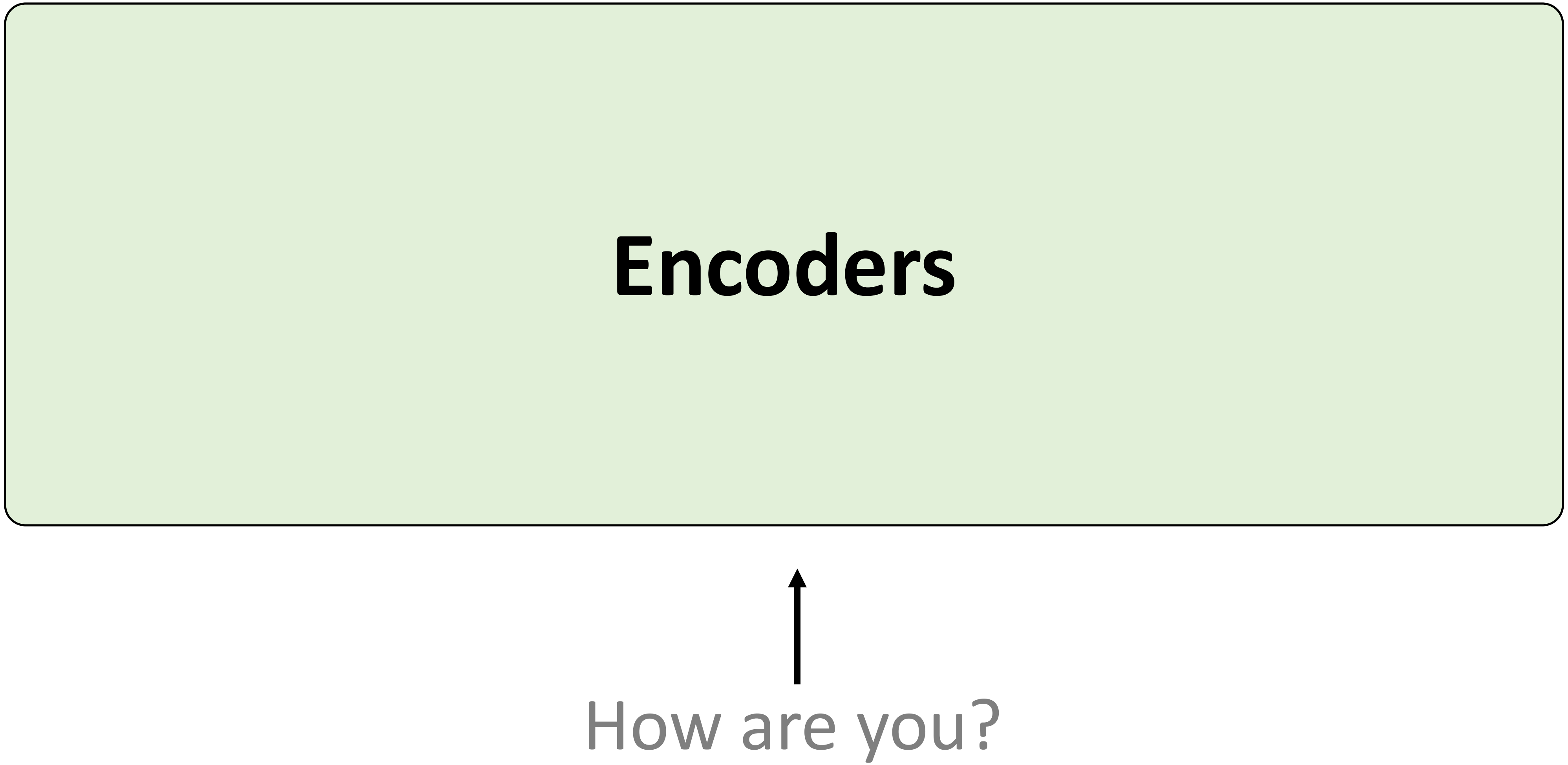
Decoders

transformer

En

How are you?

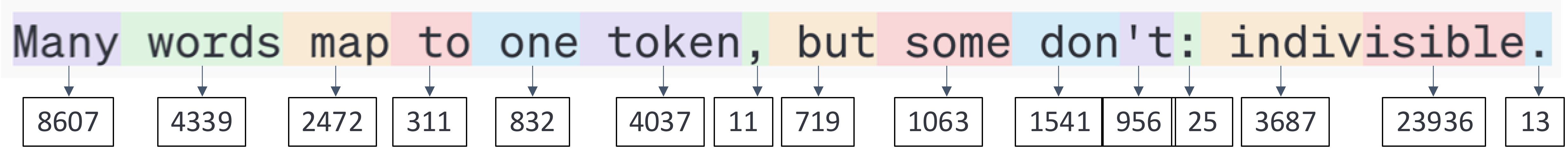
<start> 你 好 嗎 ？ <end>



Encoders

How are you?

Tokenization



Many words map to one token, but some don't: indivisible.

8607 4339 2472 311 832 4037 11 719 1063 1541 956 25 3687 23936 13

One-hot encoding

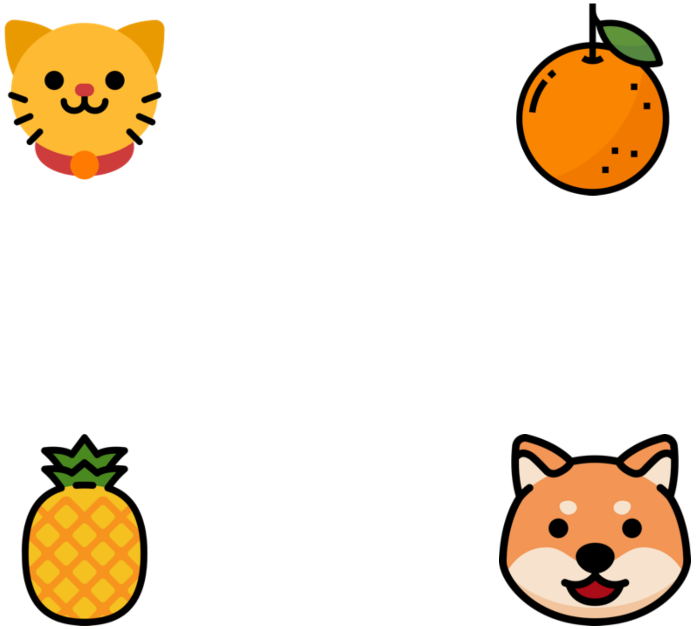
# tokens

$$\begin{array}{ccccc}
\text{cat} & \text{dog} & \text{bear} & \text{cow} & \text{indiv} \\
\begin{bmatrix}1\\0\\0\\0\\0\\\vdots\\0\end{bmatrix} &
\begin{bmatrix}0\\1\\0\\0\\0\\\vdots\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\1\\0\\0\\\vdots\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\1\\0\\\vdots\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\0\end{bmatrix}
\end{array}$$

Value 1 at 3687th entry

TOKEN EMBEDDING

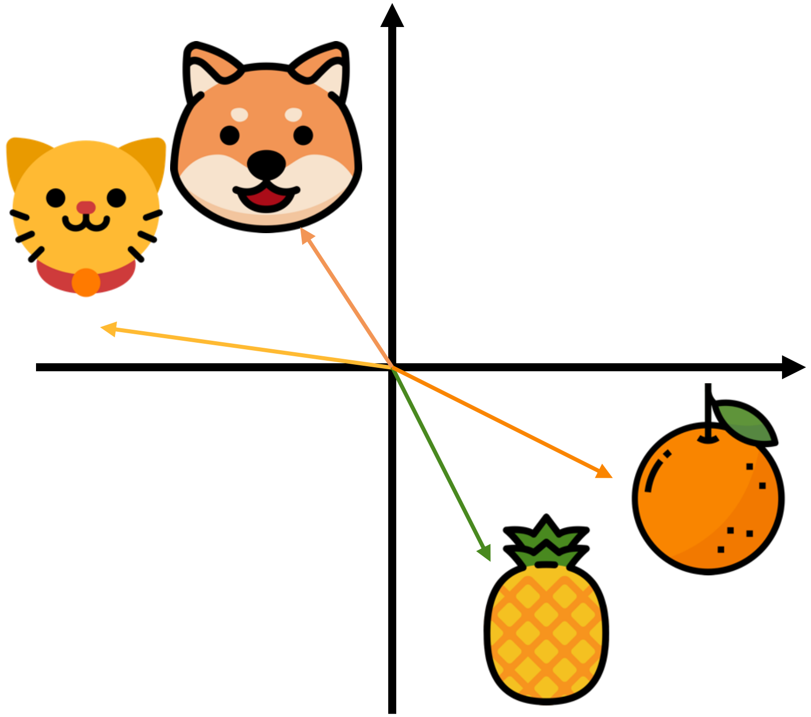
One-hot encoding



$$\begin{array}{ccccc} \text{cat} & \text{dog} & \text{bear} & \text{cow} & \text{indiv} \\ \begin{bmatrix}1\\0\\0\\0\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\1\\0\\0\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\0\\1\\0\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\0\\0\\1\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\0\end{bmatrix} \end{array}$$

Value 1 at 3687th entry

TOKEN EMBEDDING

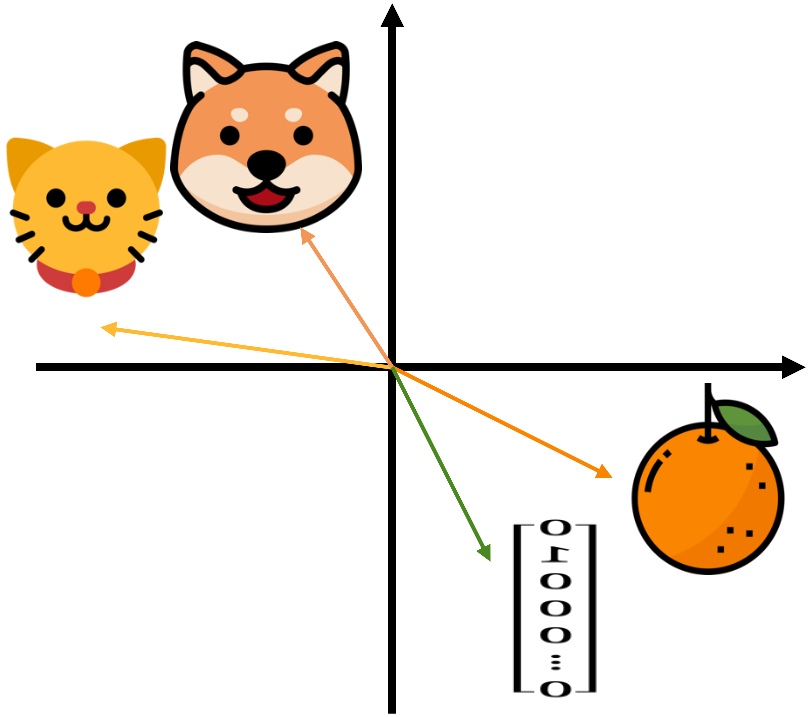


Embedding Space

$$\begin{array}{ccccc} \text{cat} & \text{dog} & \text{bear} & \text{cow} & \text{indiv} \\ \begin{bmatrix}1\\0\\0\\0\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\1\\0\\0\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\0\\1\\0\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\0\\0\\1\\0\\\vdots\\0\end{bmatrix} & \begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\0\end{bmatrix} \end{array}$$

Value 1 at 3687th entry

TOKEN EMBEDDING



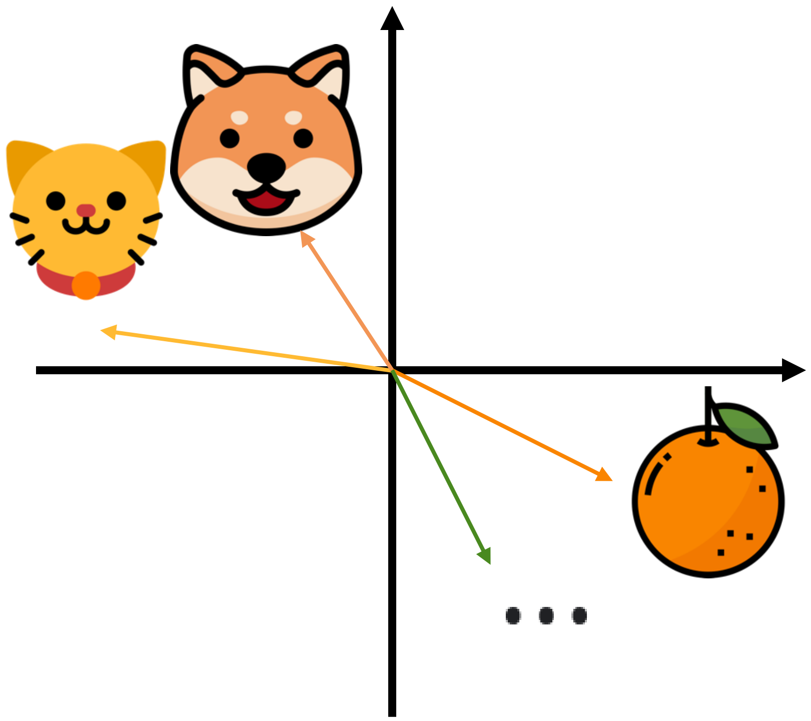
Embedding Space

$$d \begin{bmatrix}0.5\\2.7\\1.2\\\vdots\\0.2\end{bmatrix} = d \overset{\#\text{ tokens}}{\begin{bmatrix} & W_E & \end{bmatrix}} \overset{\text{dog}}{\begin{bmatrix}0\\1\\0\\0\\0\\\vdots\\0\end{bmatrix}}$$

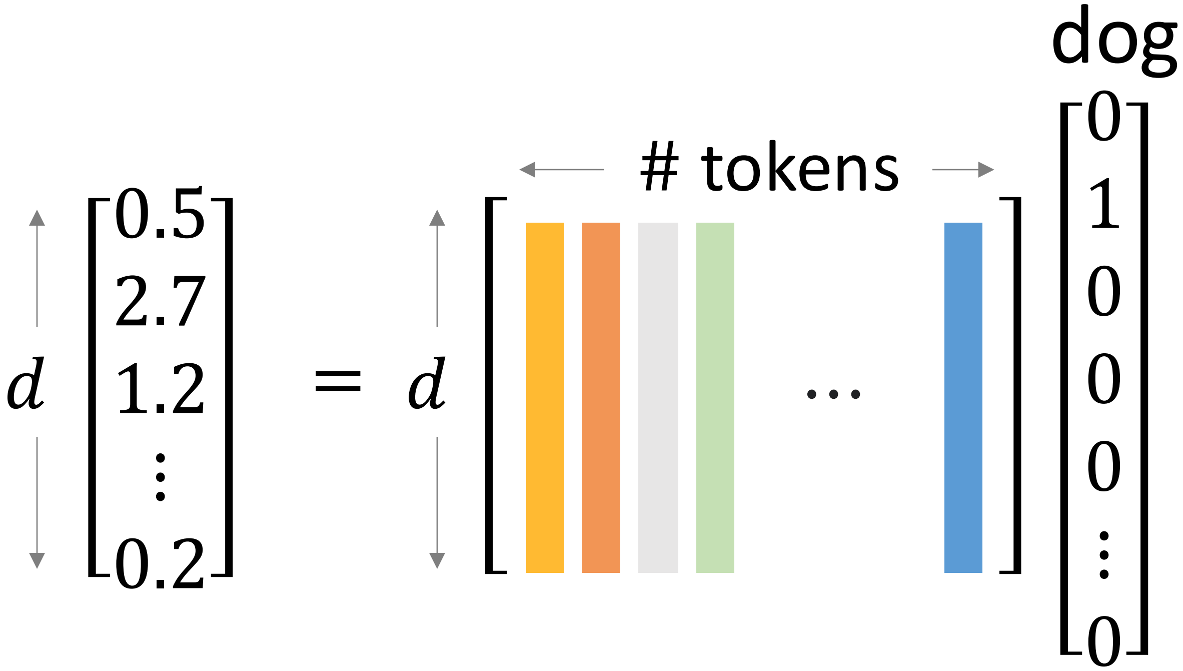
Embedded token

Embedding Matrix

TOKEN EMBEDDING



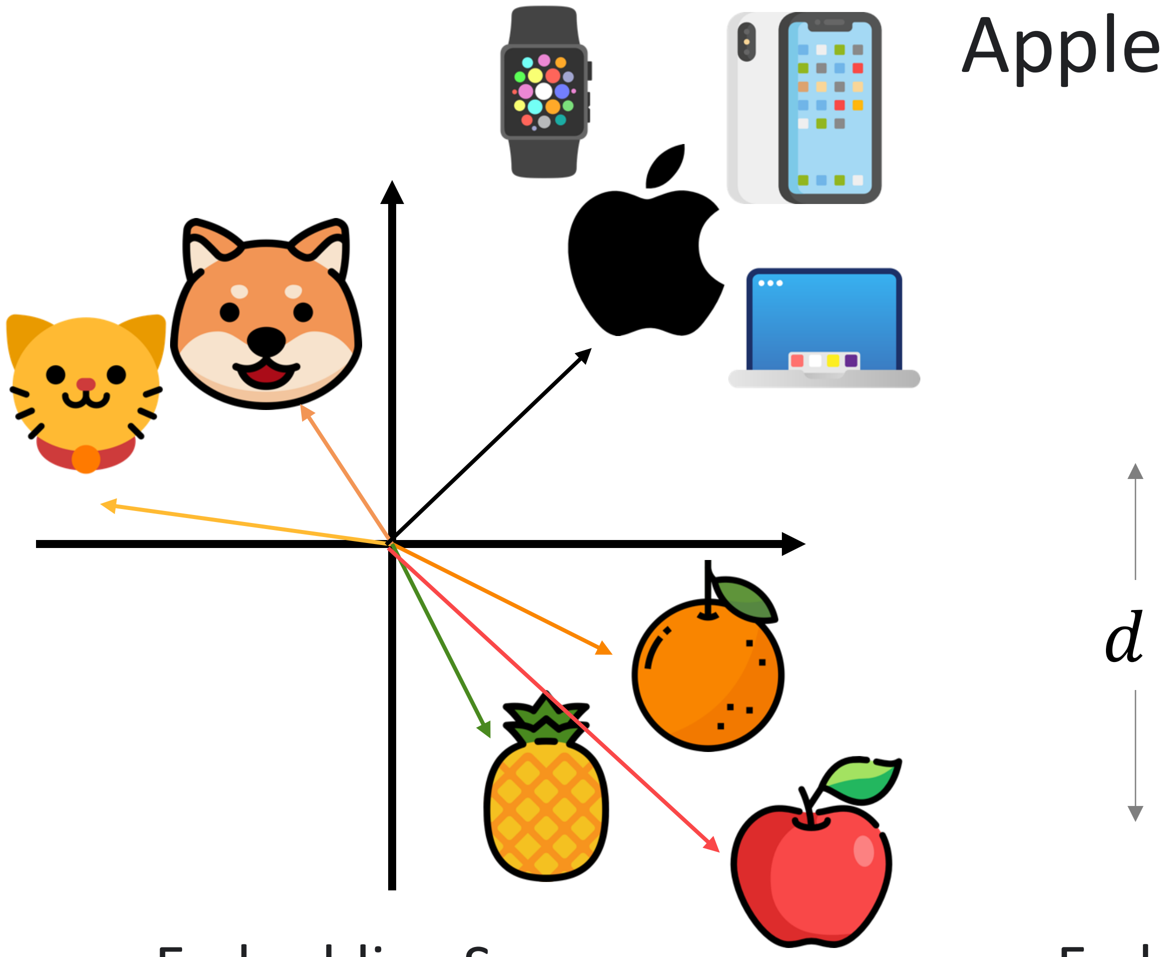
Embedding Space



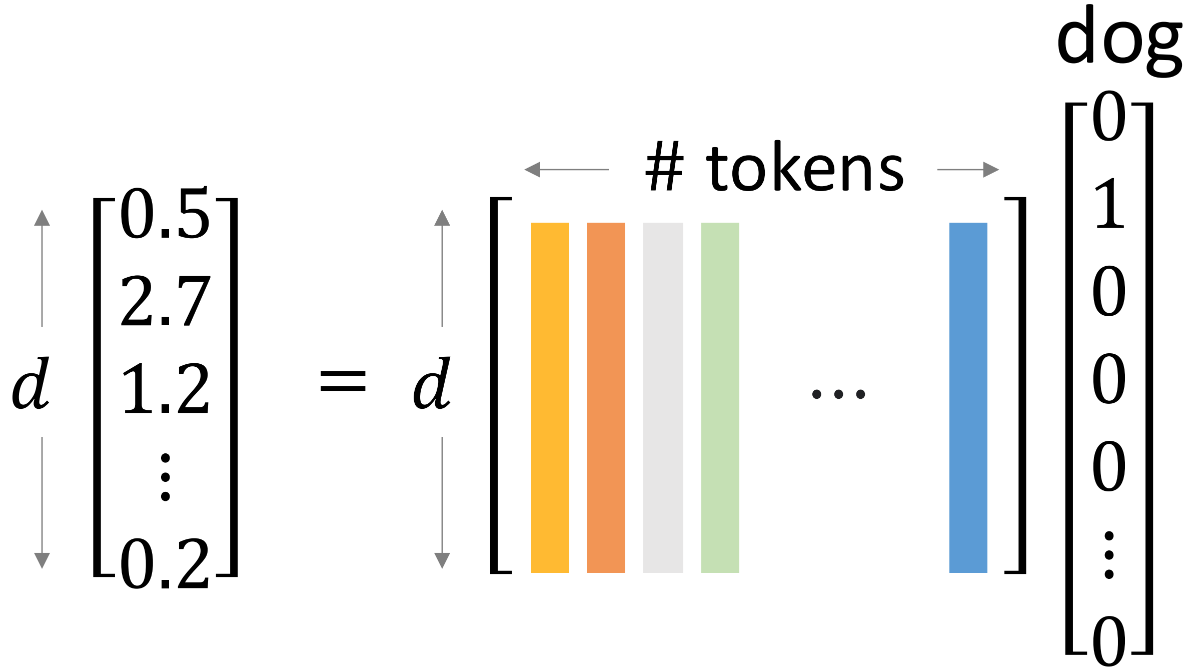
Embedded token

Embedding Matrix

TOKEN EMBEDDING



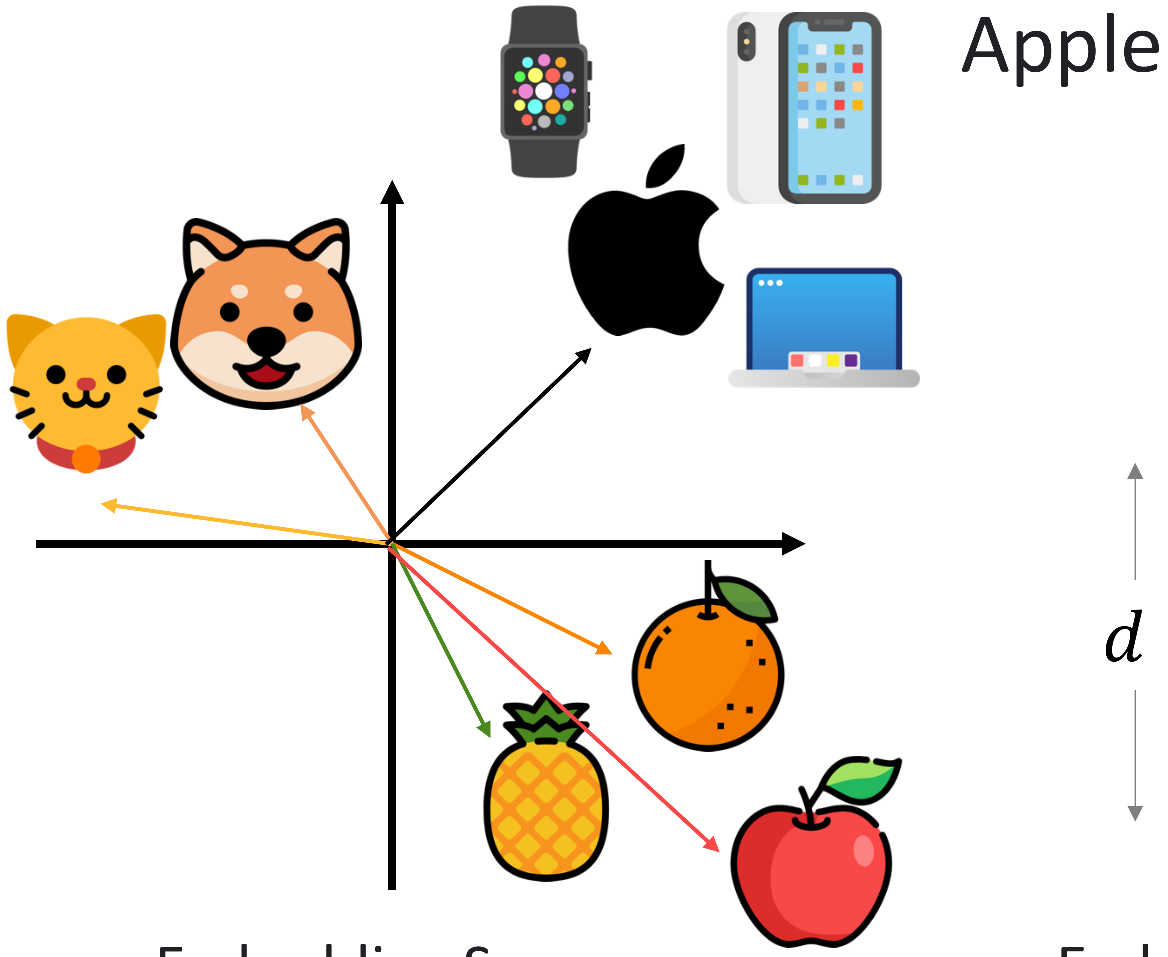
Embedding Space



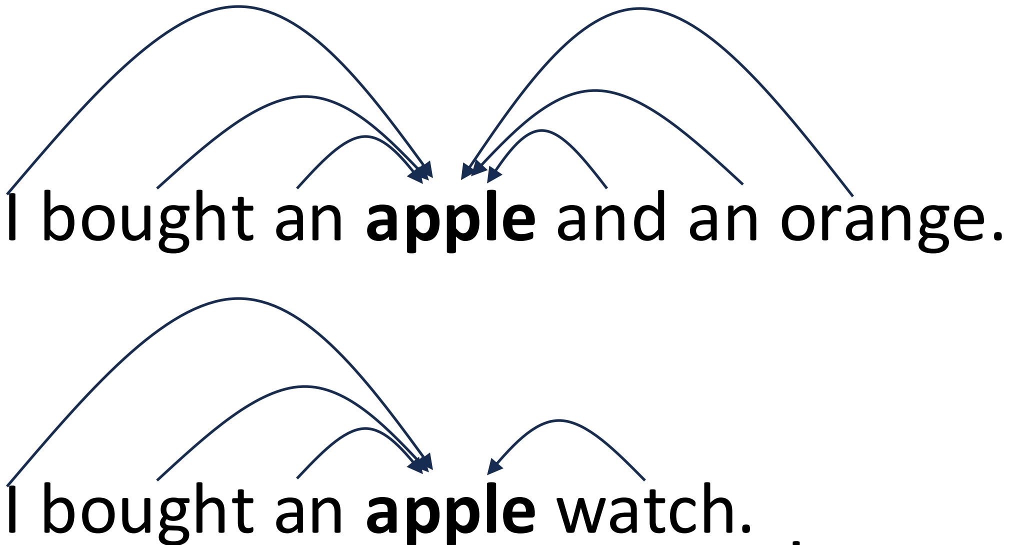
Embedded token

Embedding Matrix

TOKEN EMBEDDING

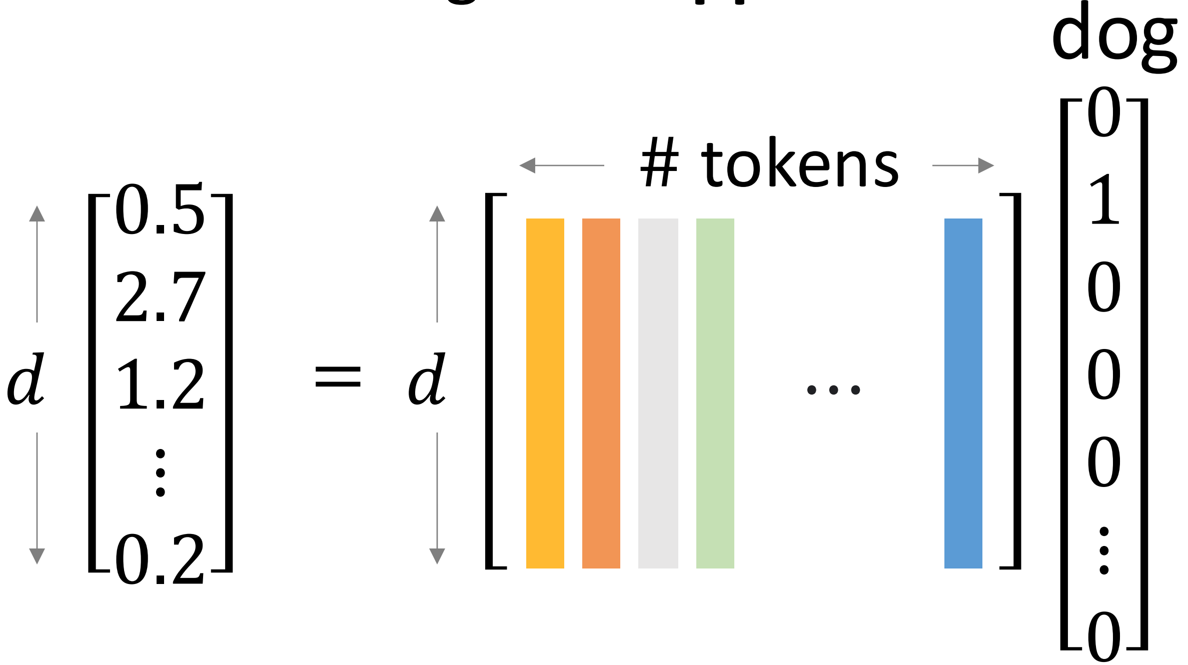


Embedding Space



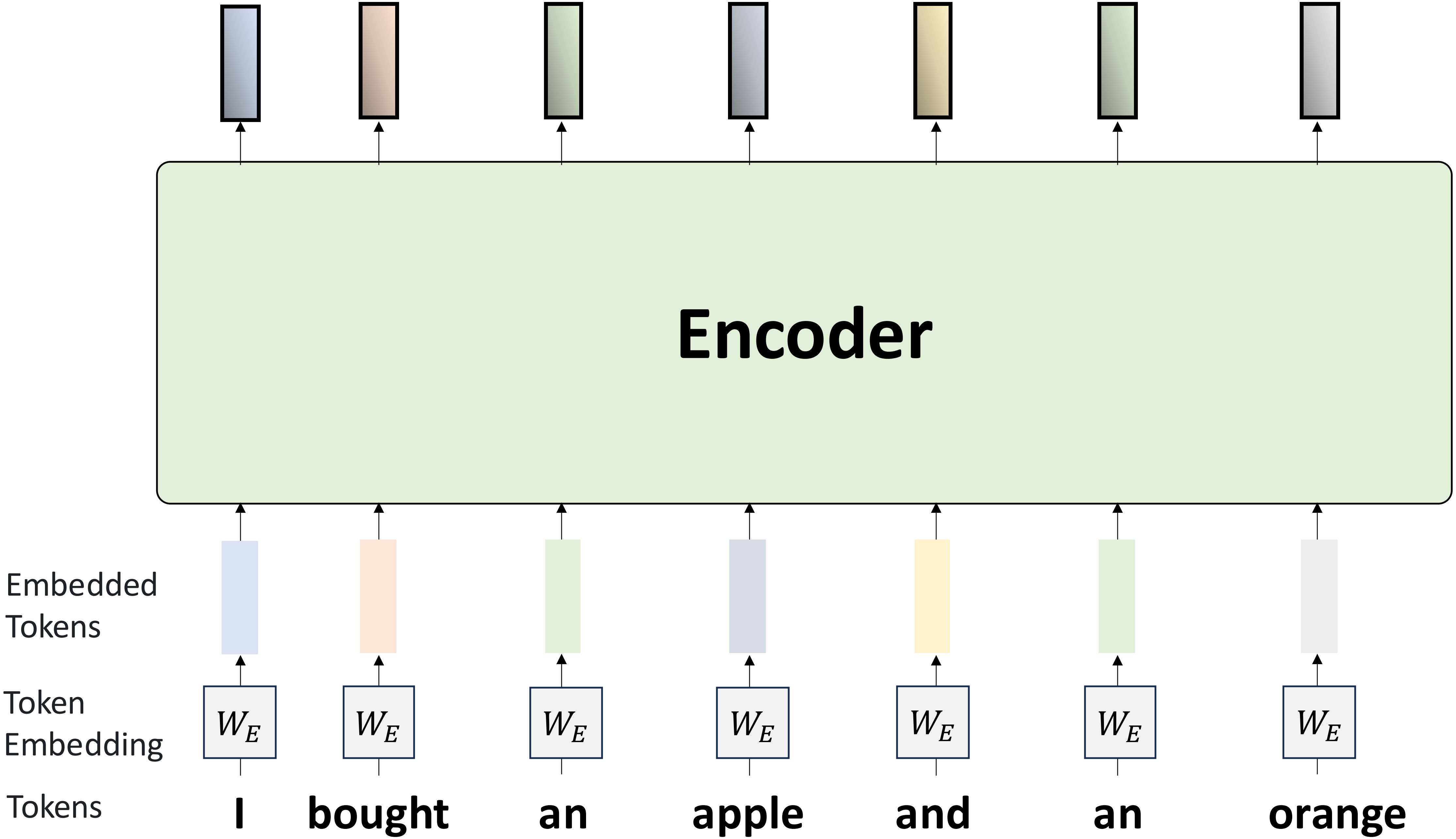
I bought an **apple** and an orange.

I bought an **apple** watch.



Embedded token

Embedding Matrix



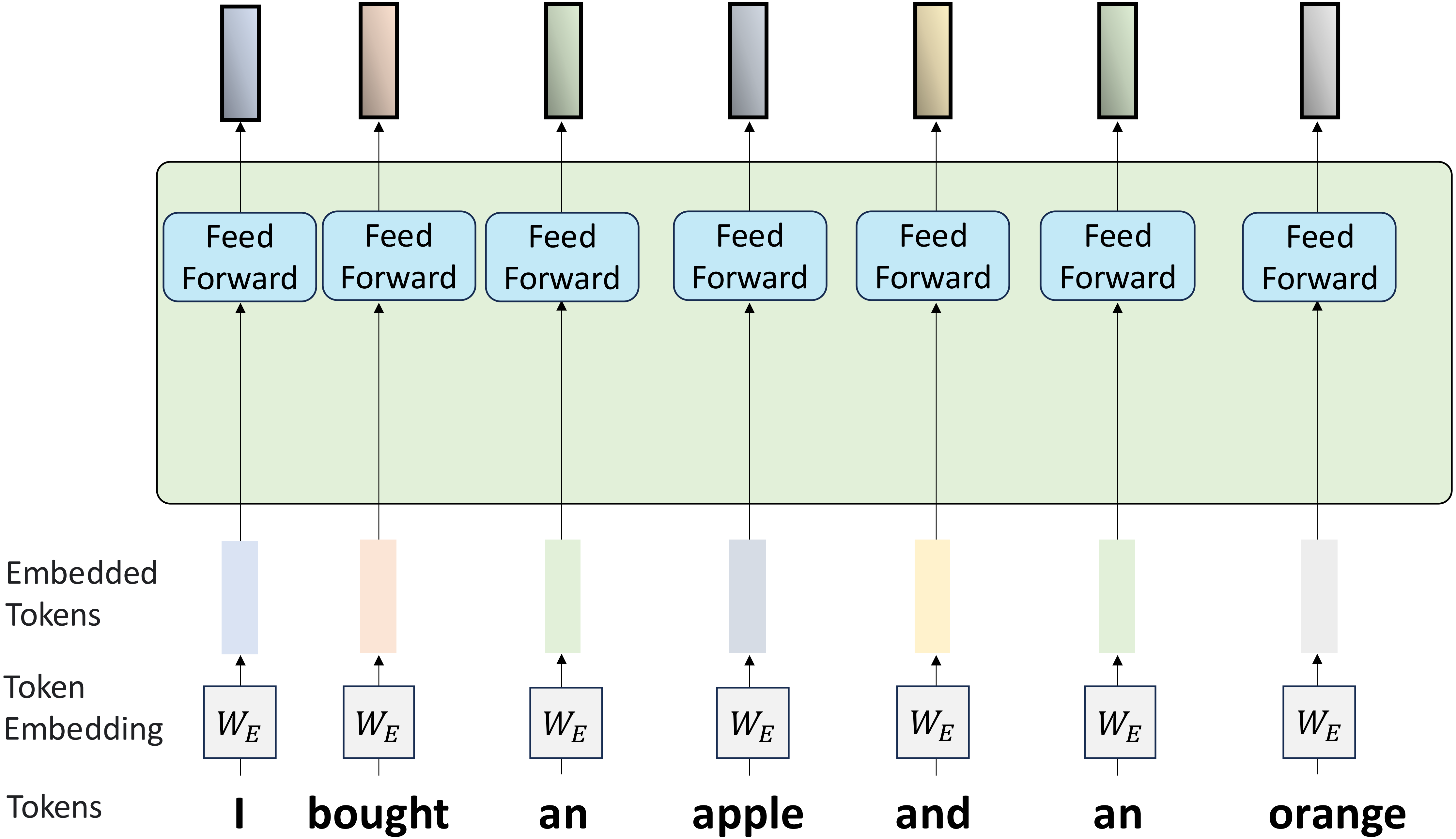
Encoder

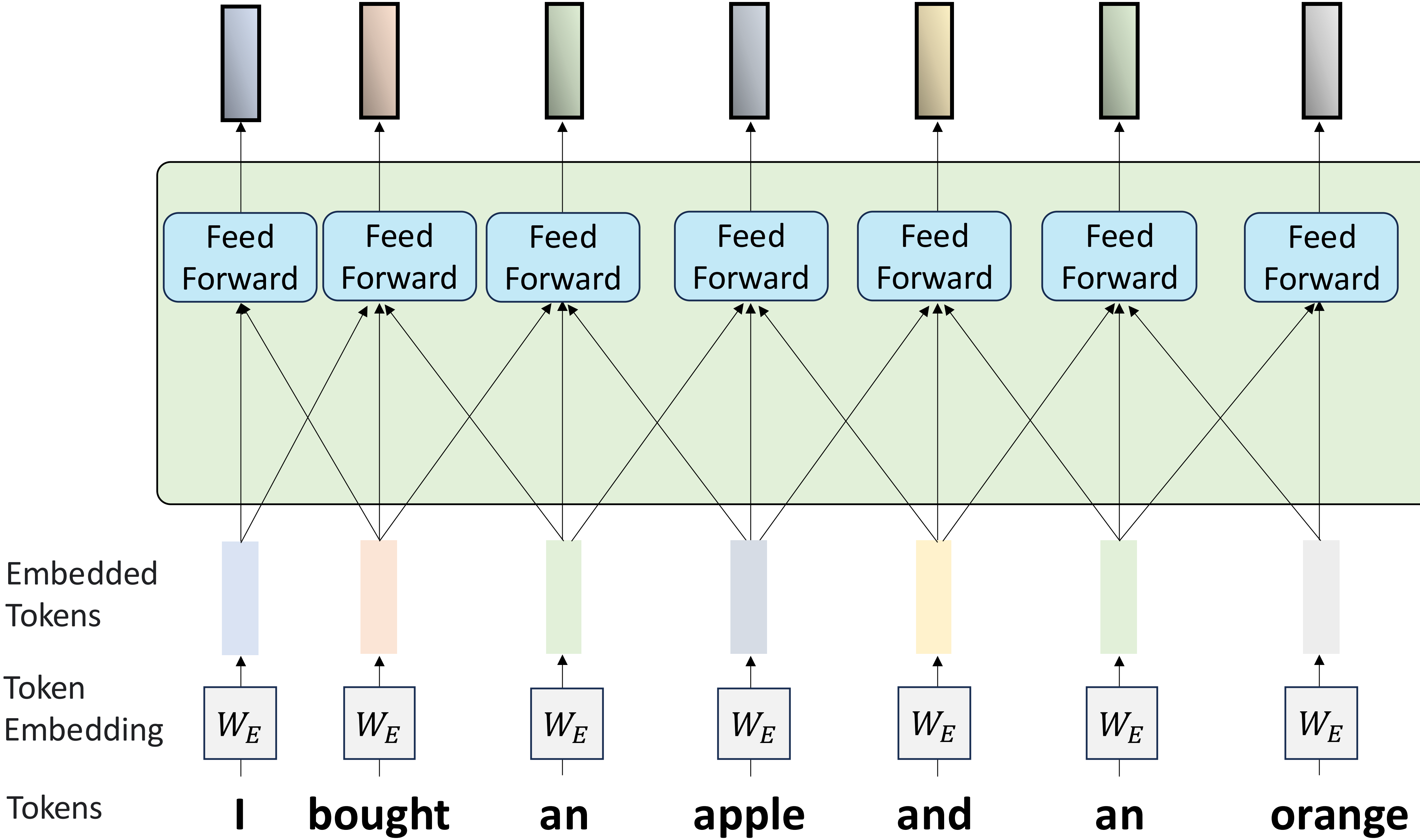
Embedded Tokens

Token Embedding

Tokens

I bought an apple and an orange


Feed Forward
Feed Forward
Feed Forward
Feed Forward
Feed Forward
Feed Forward
Feed Forward
Embedded Tokens
Token Embedding
W_E
W_E
W_E
W_E
W_E
W_E
W_E
Tokens
I
bought
an
apple
and
an
orange



Feed Forward

Feed Forward

Feed Forward

Feed Forward

Feed Forward

Feed Forward

Feed Forward

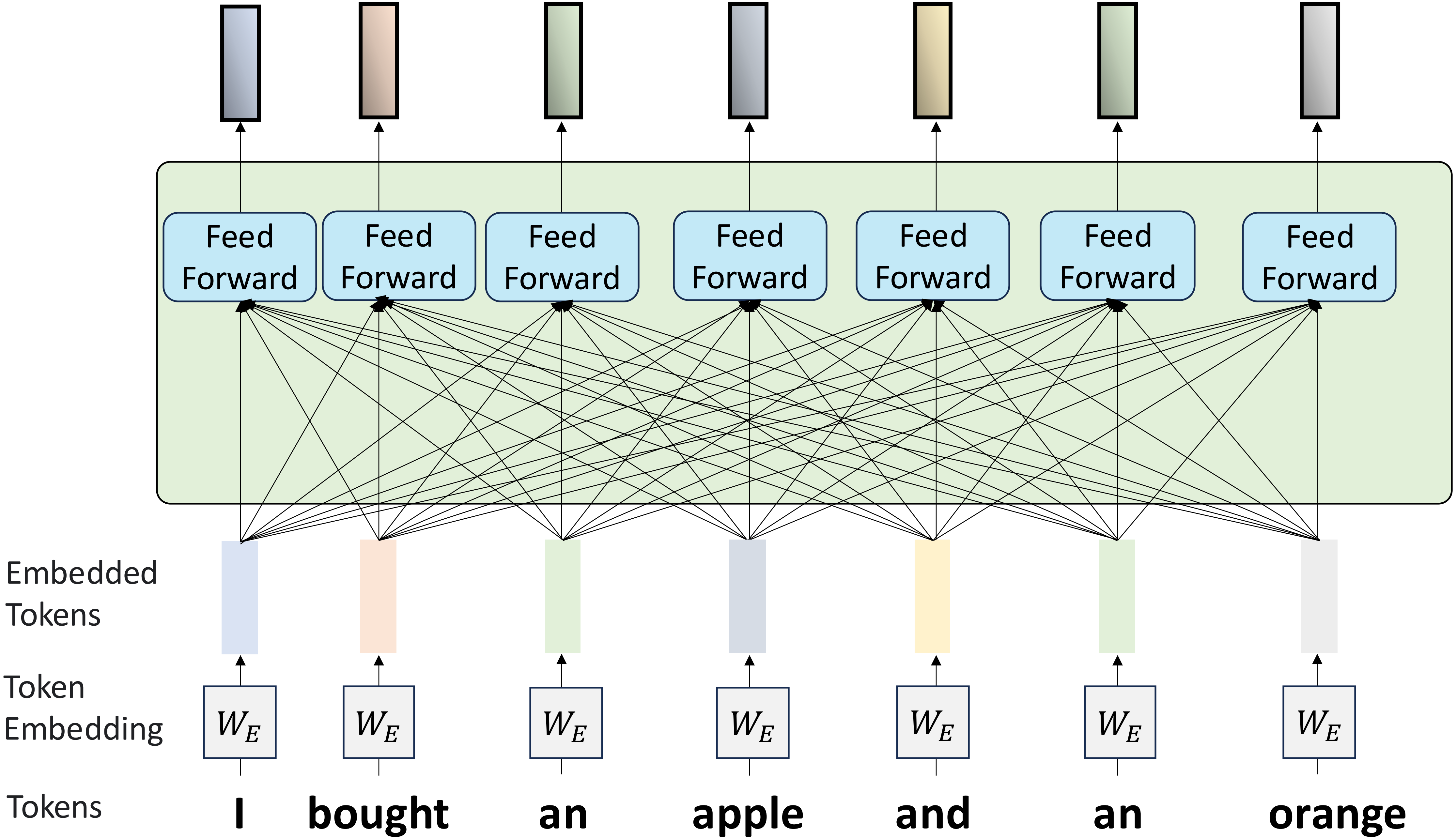
Embedded Tokens

Token Embedding

W_E W_E W_E W_E W_E W_E W_E

Tokens

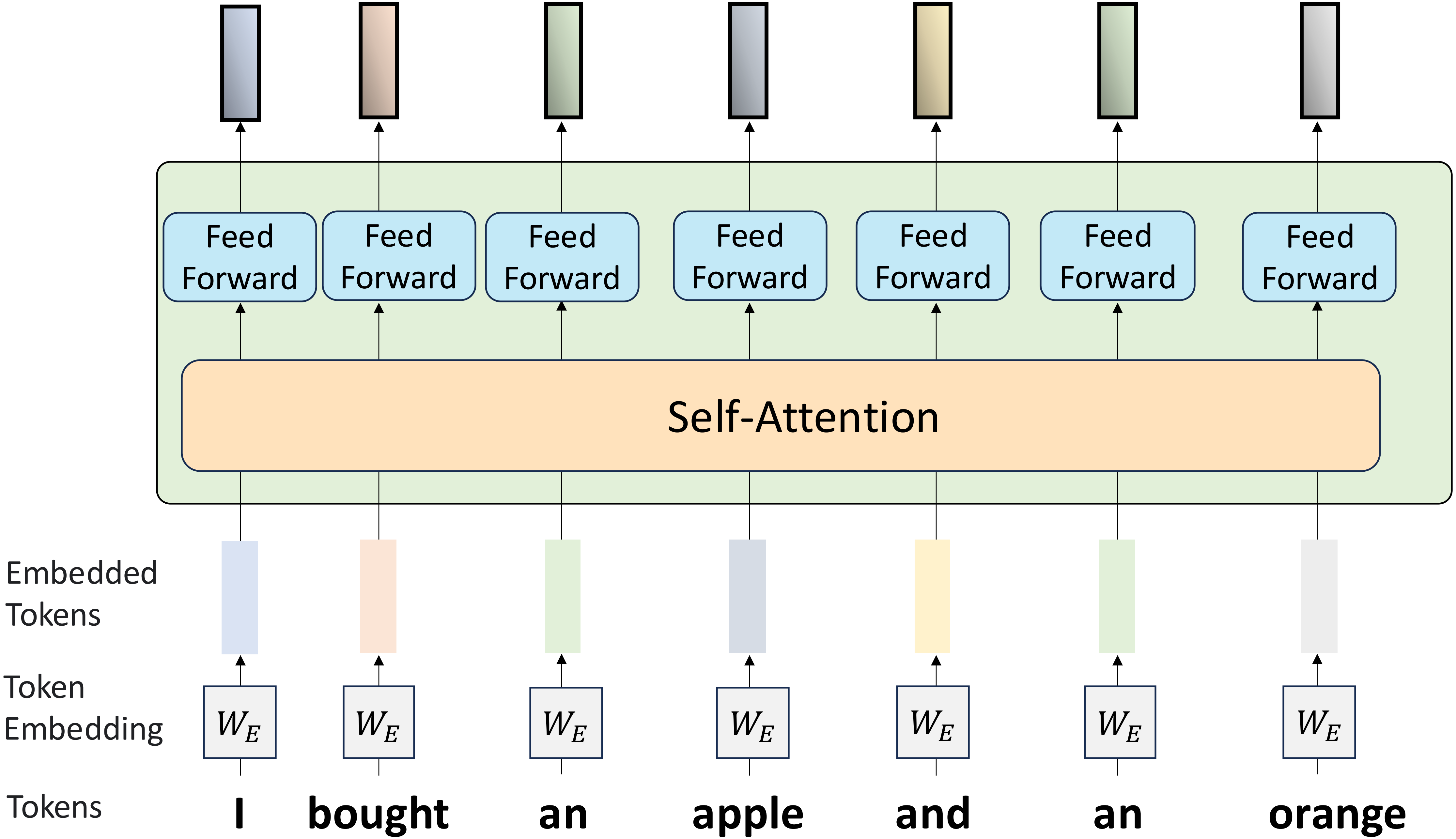
I bought an apple and an orange



Embedded Tokens

Token Embedding

Tokens



Feed Forward | Feed Forward | Feed Forward | Feed Forward | Feed Forward | Feed Forward | Feed Forward

Self-Attention

Embedded Tokens

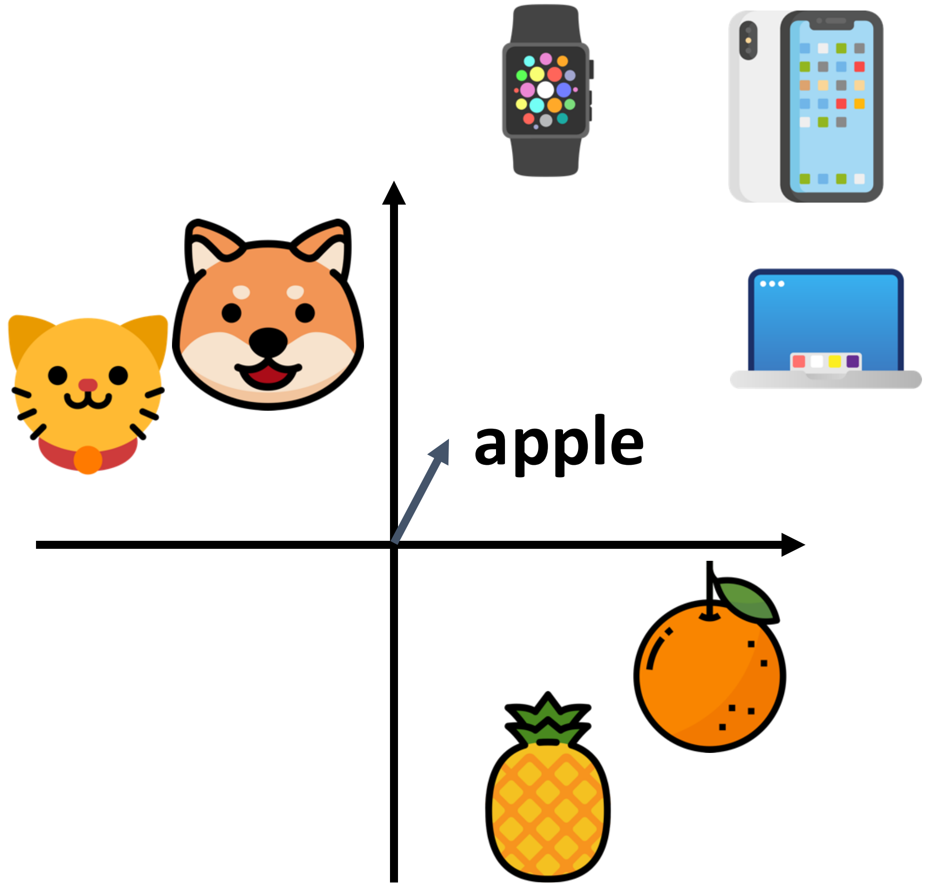
Token Embedding

W_E W_E W_E W_E W_E W_E W_E

Tokens

I bought an apple and an orange

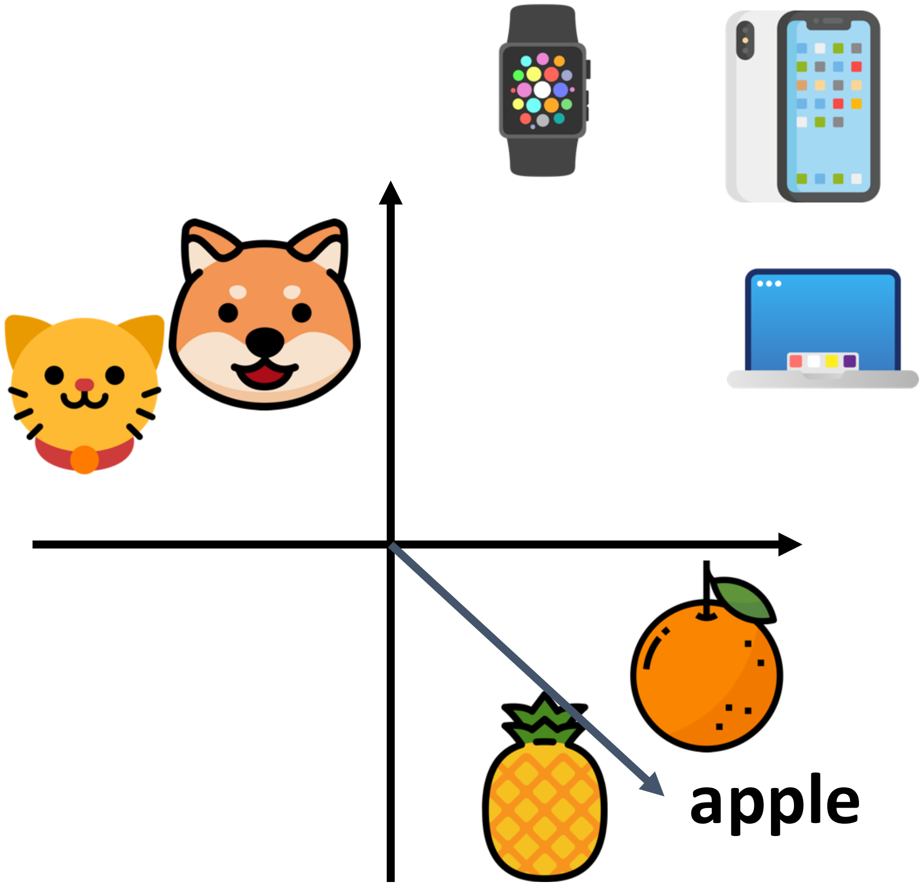
Self-Attention



Embedding Space

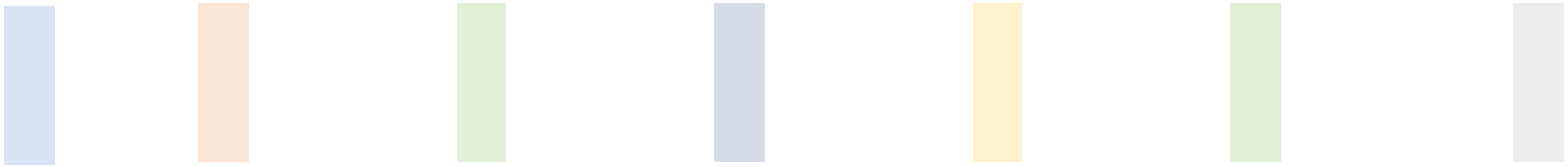
Embedded Tokens	🟦	🟧	🟩	⬜	🟨	🟩	⬜
Tokens	**I**	**bought**	**an**	**apple**	**and**	**an**	**orange**

Self-Attention



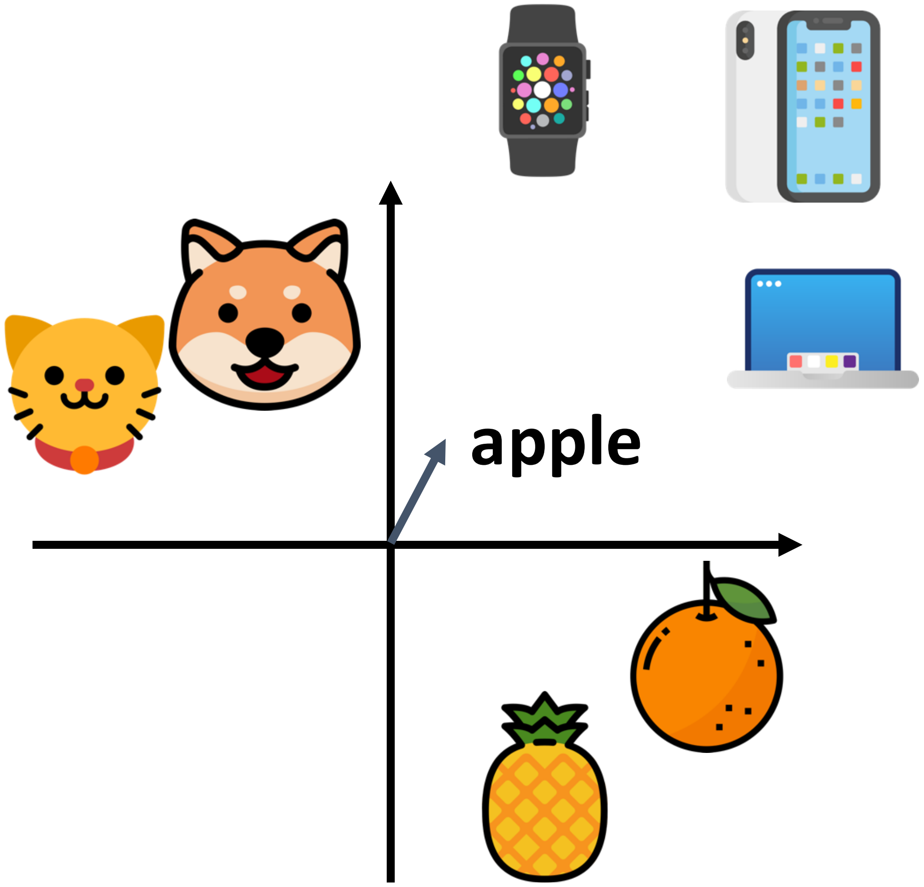
Embedding Space

Embedded Tokens

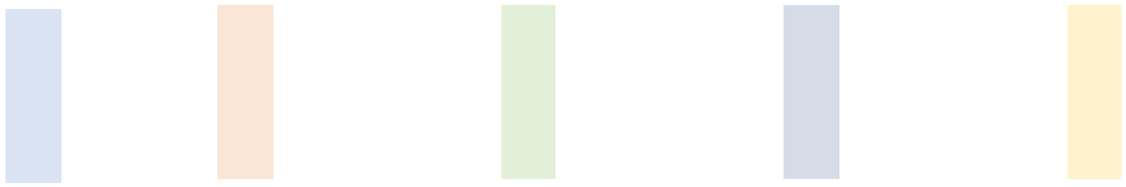


Tokens **I** **bought** **an** **apple** **and** **an** **orange**

Self-Attention



Embedding Space

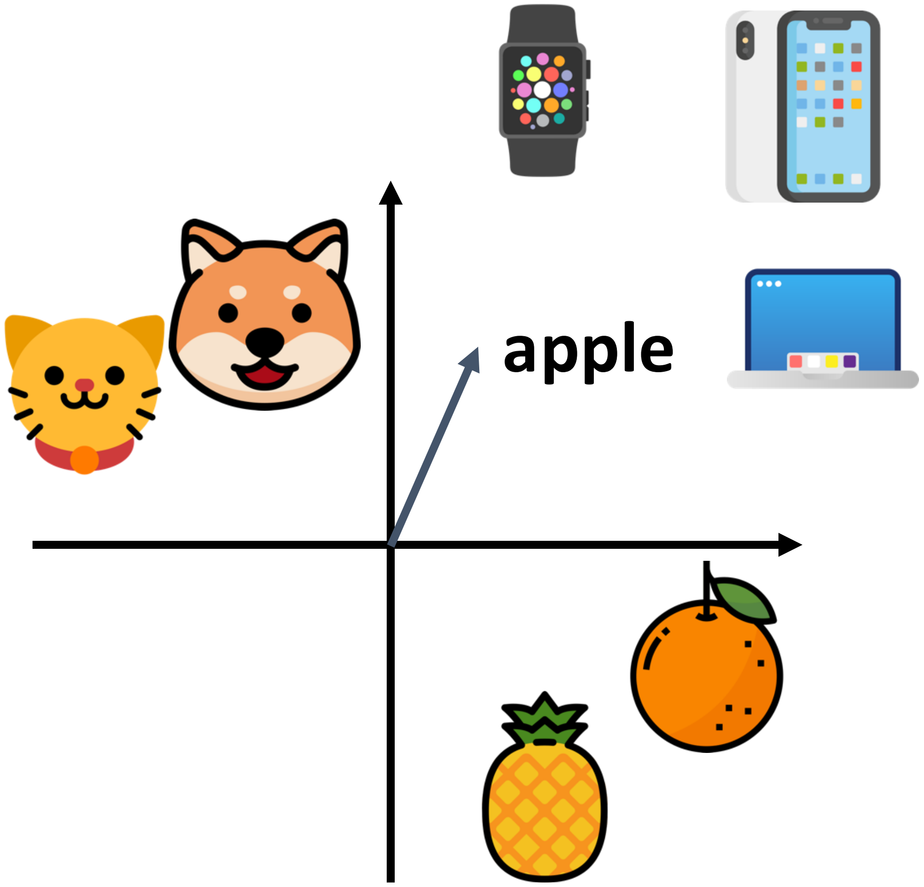


Embedded Tokens

Tokens

I **bought** **an** **apple** **watch**

Self-Attention

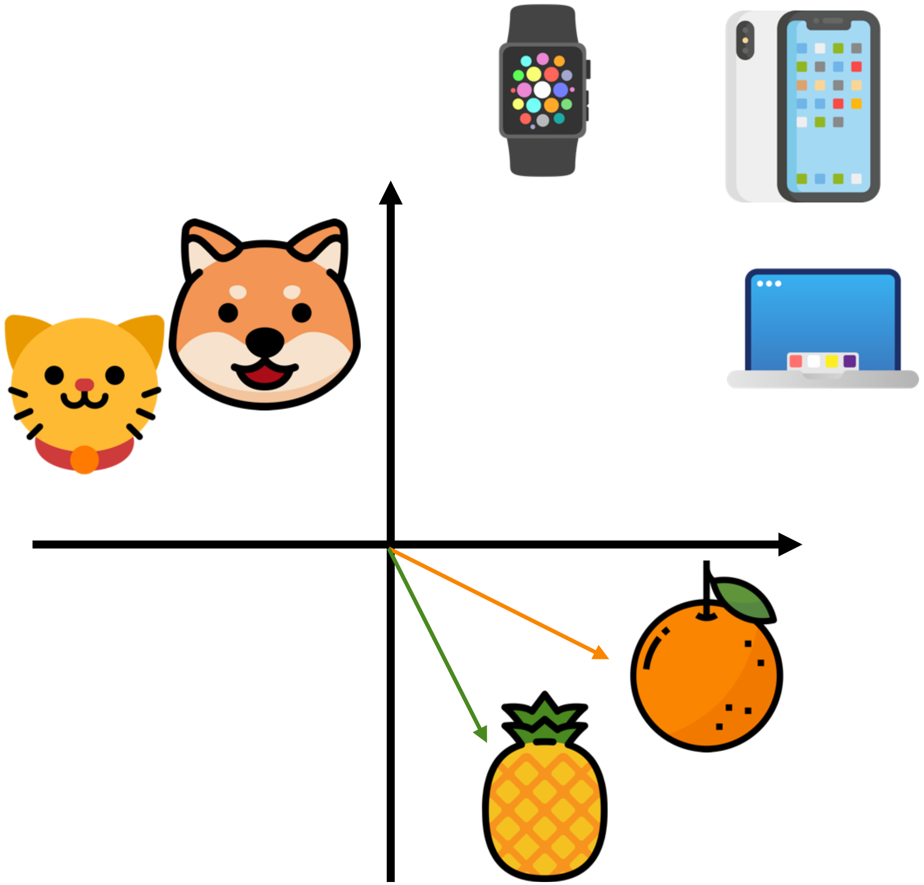


Embedding Space

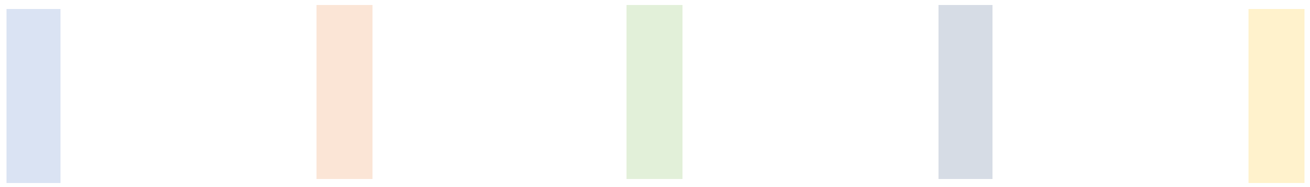
Embedded Tokens

Tokens **I** **bought** **an** **apple** **watch**

Self-Attention



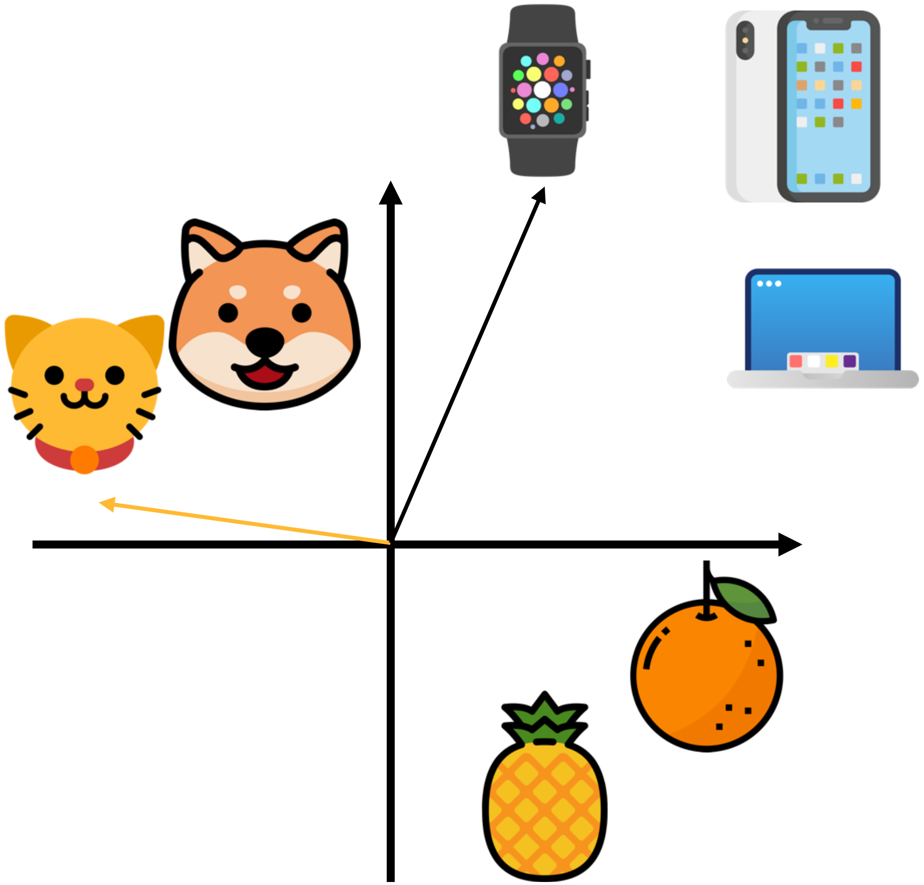
Embedding Space



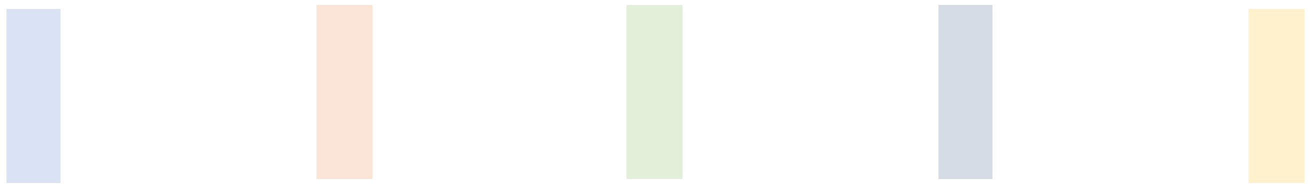
Embedded Tokens

Tokens **I** **bought** **an** **apple** **watch**

Self-Attention



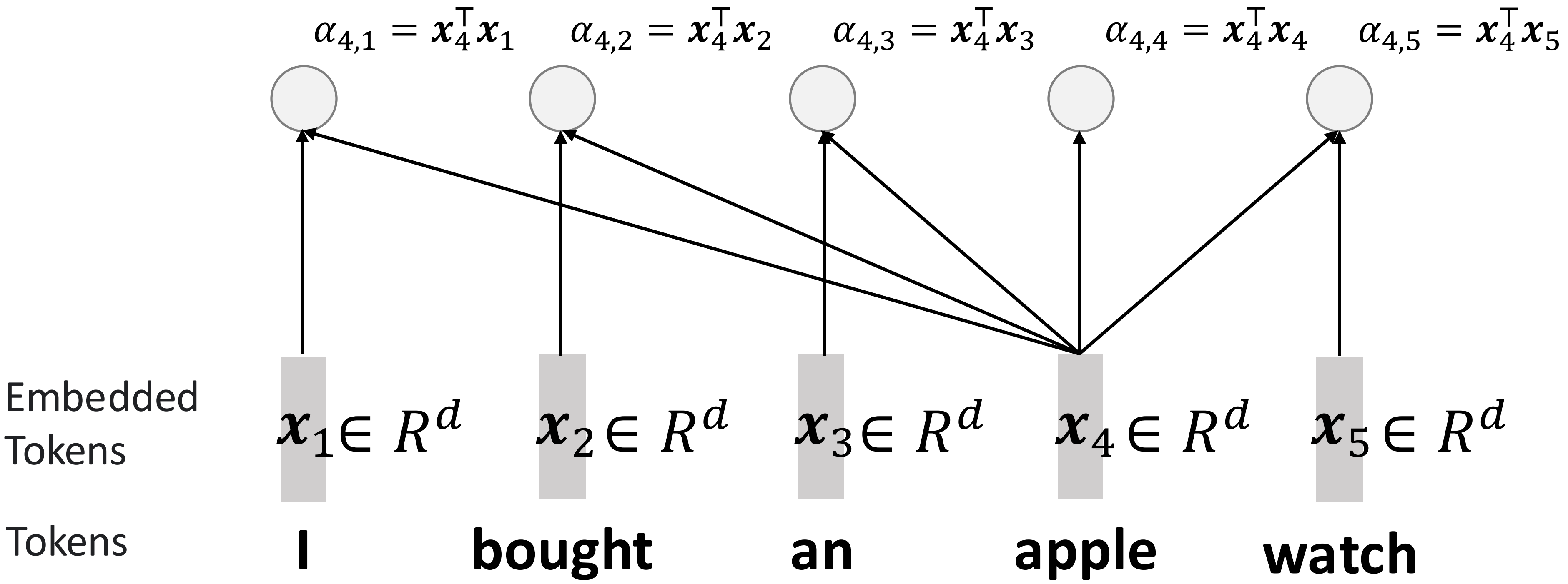
Embedding Space

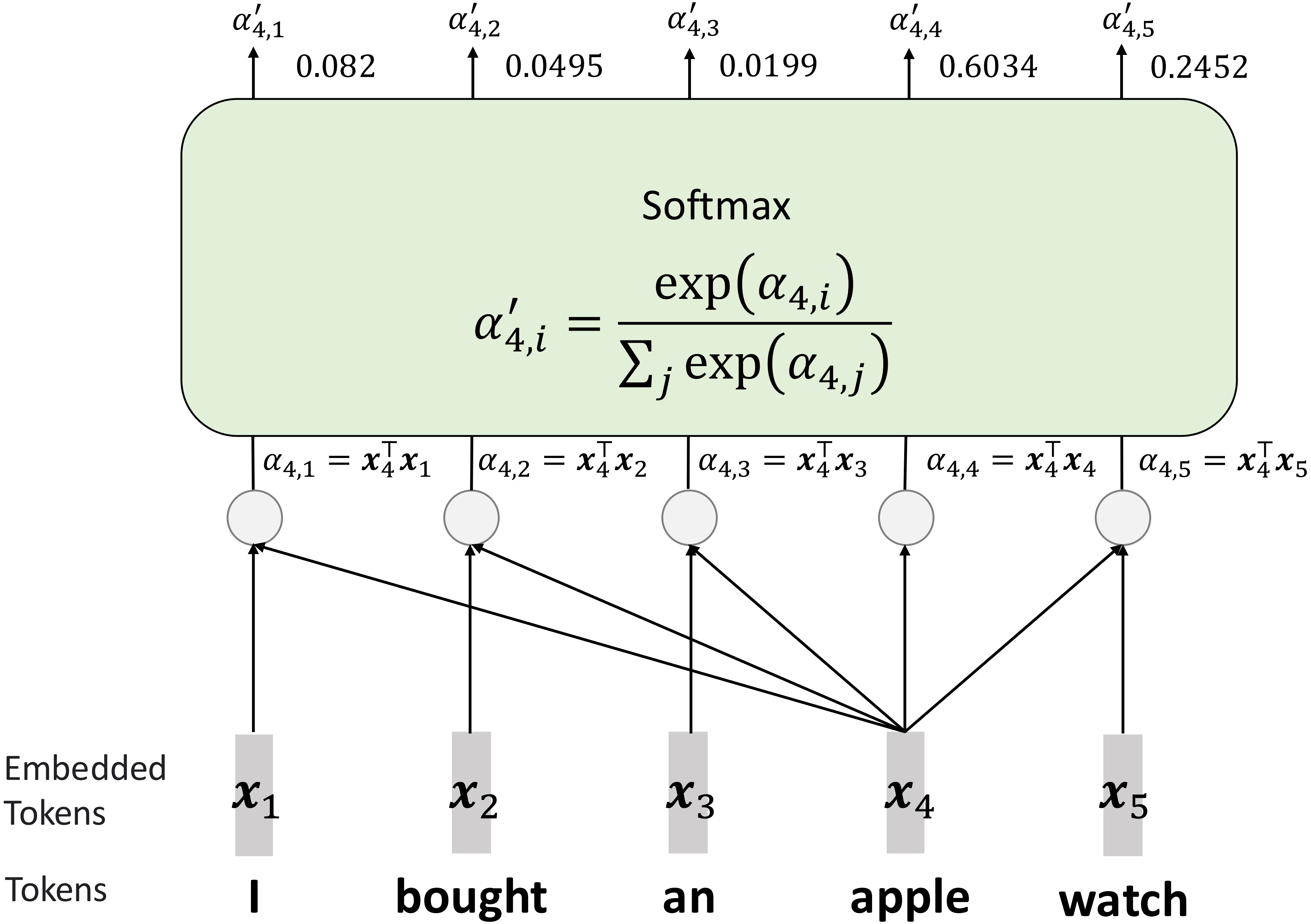


Embedded Tokens

Tokens **I** **bought** **an** **apple** **watch**

Self-Attention


$\alpha_{4,1} = \boldsymbol{x}_4^\top \boldsymbol{x}_1$
$\alpha_{4,2} = \boldsymbol{x}_4^\top \boldsymbol{x}_2$
$\alpha_{4,3} = \boldsymbol{x}_4^\top \boldsymbol{x}_3$
$\alpha_{4,4} = \boldsymbol{x}_4^\top \boldsymbol{x}_4$
$\alpha_{4,5} = \boldsymbol{x}_4^\top \boldsymbol{x}_5$
Embedded Tokens
$\boldsymbol{x}_1 \in R^d$
$\boldsymbol{x}_2 \in R^d$
$\boldsymbol{x}_3 \in R^d$
$\boldsymbol{x}_4 \in R^d$
$\boldsymbol{x}_5 \in R^d$
Tokens
I
bought
an
apple
watch



$\alpha'_{4,1}$ 0.082

$\alpha'_{4,2}$ 0.0495

$\alpha'_{4,3}$ 0.0199

$\alpha'_{4,4}$ 0.6034

$\alpha'_{4,5}$ 0.2452

Softmax

$$\alpha'_{4,i} = \frac{\exp(\alpha_{4,i})}{\sum_j \exp(\alpha_{4,j})}$$

$\alpha_{4,1} = \boldsymbol{x}_4^\top \boldsymbol{x}_1$

$\alpha_{4,2} = \boldsymbol{x}_4^\top \boldsymbol{x}_2$

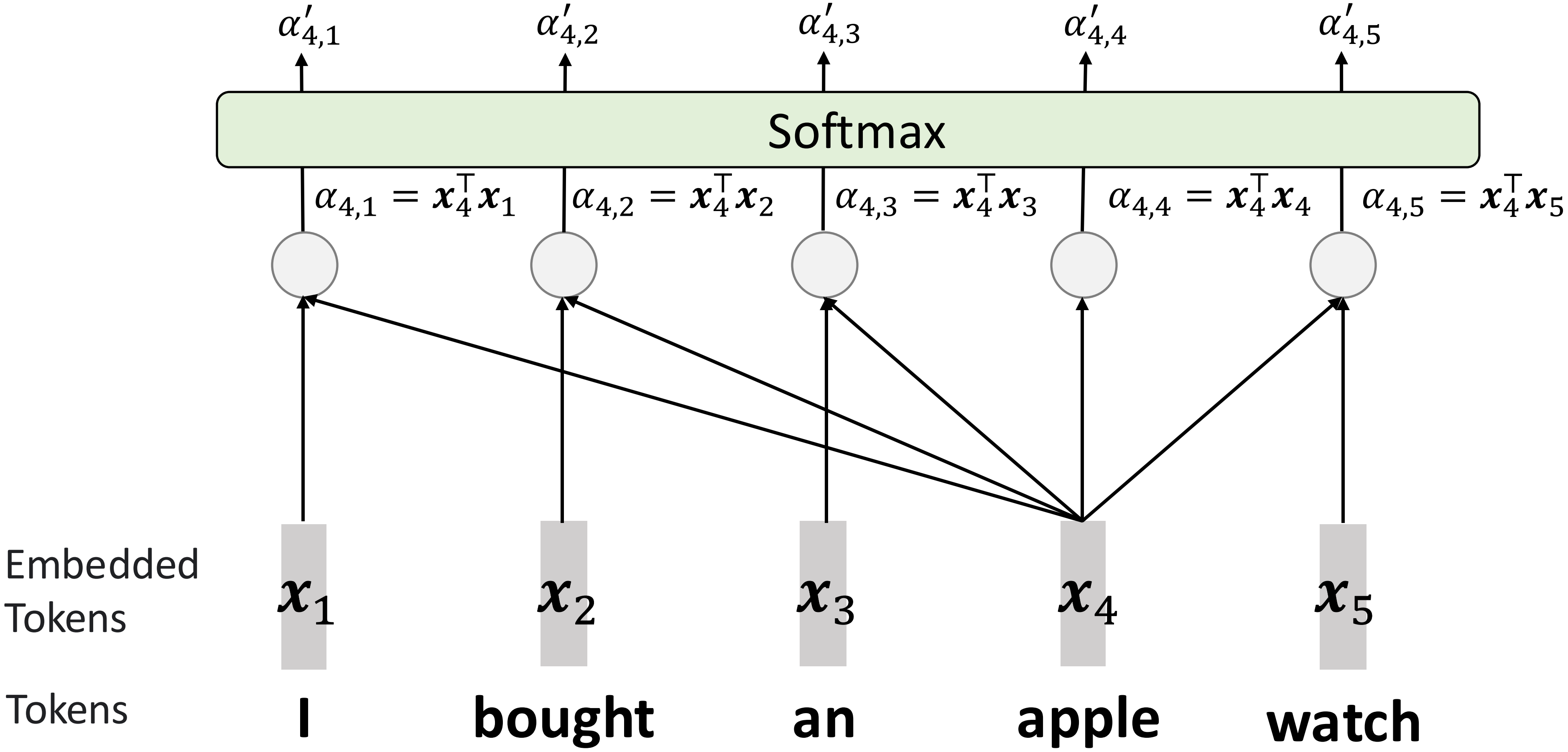
$\alpha_{4,3} = \boldsymbol{x}_4^\top \boldsymbol{x}_3$

$\alpha_{4,4} = \boldsymbol{x}_4^\top \boldsymbol{x}_4$

$\alpha_{4,5} = \boldsymbol{x}_4^\top \boldsymbol{x}_5$

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens **I** **bought** **an** **apple** **watch**



$\alpha'_{4,1}$ $\alpha'_{4,2}$ $\alpha'_{4,3}$ $\alpha'_{4,4}$ $\alpha'_{4,5}$

Softmax

$\alpha_{4,1} = \boldsymbol{x}_4^\top \boldsymbol{x}_1$

$\alpha_{4,2} = \boldsymbol{x}_4^\top \boldsymbol{x}_2$

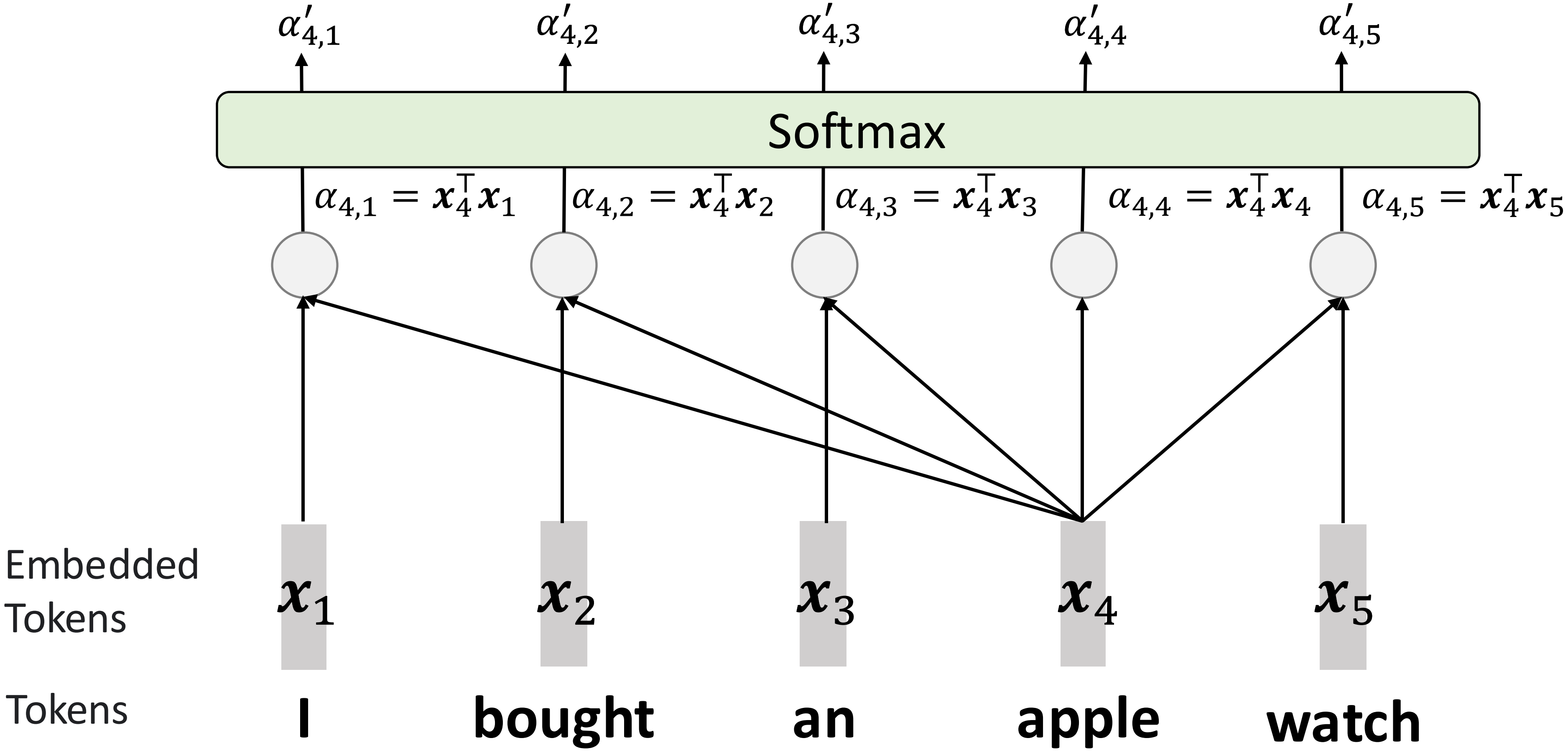
$\alpha_{4,3} = \boldsymbol{x}_4^\top \boldsymbol{x}_3$

$\alpha_{4,4} = \boldsymbol{x}_4^\top \boldsymbol{x}_4$

$\alpha_{4,5} = \boldsymbol{x}_4^\top \boldsymbol{x}_5$

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens **I** **bought** **an** **apple** **watch**



$\alpha'_{4,1}$ $\alpha'_{4,2}$ $\alpha'_{4,3}$ $\alpha'_{4,4}$ $\alpha'_{4,5}$

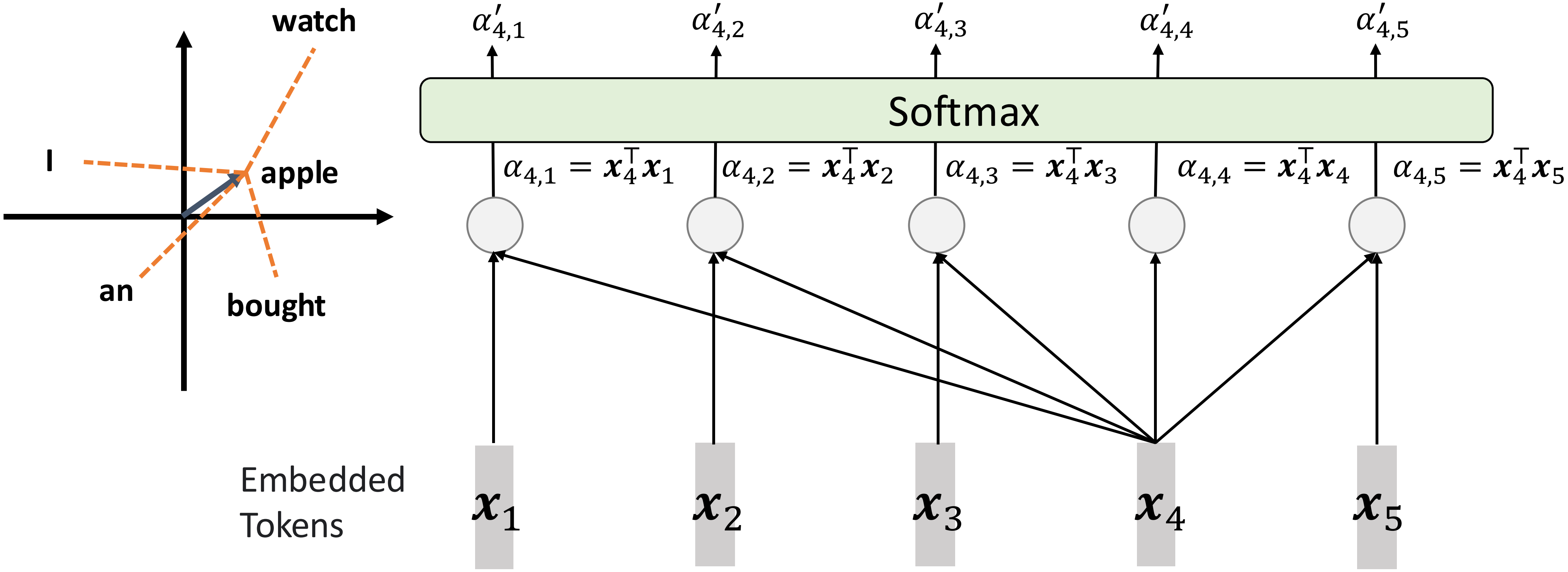
Softmax

$\alpha_{4,1} = \boldsymbol{x}_4^\top \boldsymbol{x}_1$ $\alpha_{4,2} = \boldsymbol{x}_4^\top \boldsymbol{x}_2$ $\alpha_{4,3} = \boldsymbol{x}_4^\top \boldsymbol{x}_3$ $\alpha_{4,4} = \boldsymbol{x}_4^\top \boldsymbol{x}_4$ $\alpha_{4,5} = \boldsymbol{x}_4^\top \boldsymbol{x}_5$

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens **I** **bought** **an** **apple** **watch**

Updated feature $\boldsymbol{x}'_4 = \alpha'_{4,1}\boldsymbol{x}_1 + \alpha'_{4,2}\boldsymbol{x}_2 + \alpha'_{4,3}\boldsymbol{x}_3 + \alpha'_{4,4}\boldsymbol{x}_4 + \alpha'_{4,5}\boldsymbol{x}_5$



$\alpha'_{4,1}$ $\alpha'_{4,2}$ $\alpha'_{4,3}$ $\alpha'_{4,4}$ $\alpha'_{4,5}$

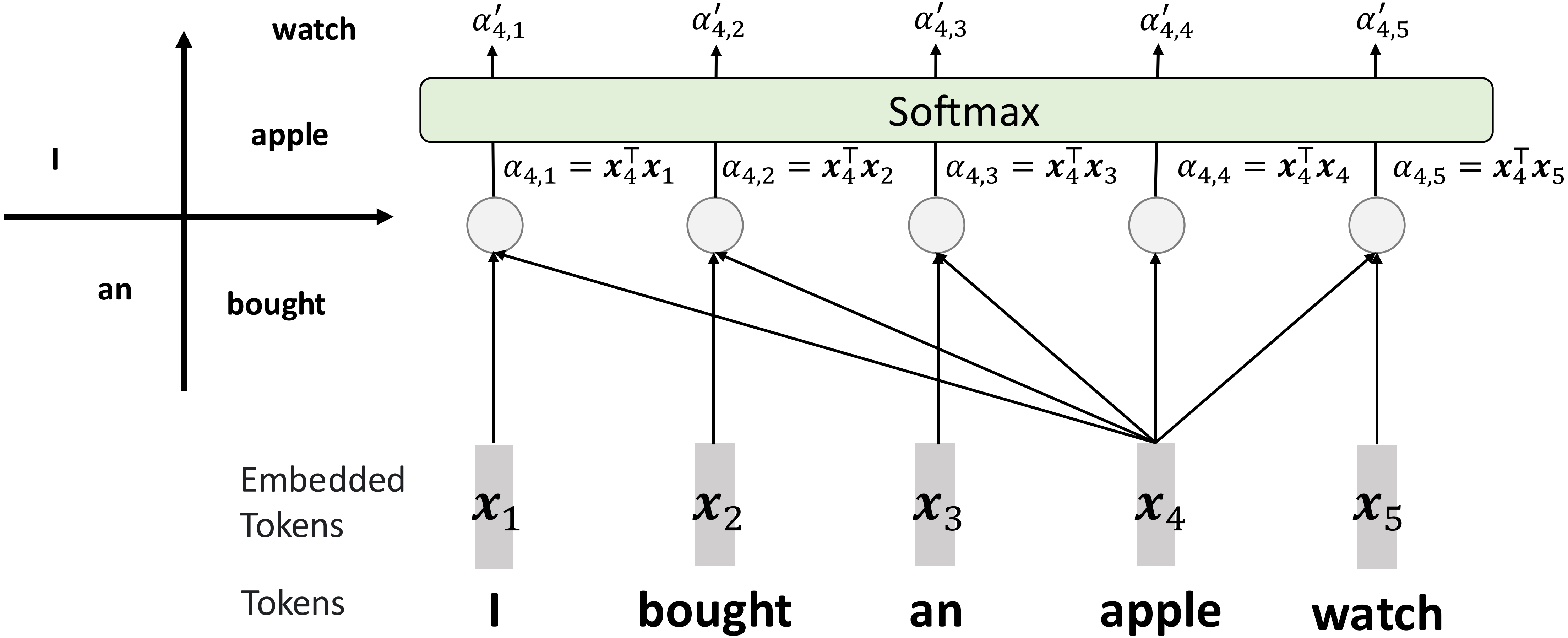
Softmax

$\alpha_{4,1} = \boldsymbol{x}_4^\top \boldsymbol{x}_1$ $\alpha_{4,2} = \boldsymbol{x}_4^\top \boldsymbol{x}_2$ $\alpha_{4,3} = \boldsymbol{x}_4^\top \boldsymbol{x}_3$ $\alpha_{4,4} = \boldsymbol{x}_4^\top \boldsymbol{x}_4$ $\alpha_{4,5} = \boldsymbol{x}_4^\top \boldsymbol{x}_5$

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

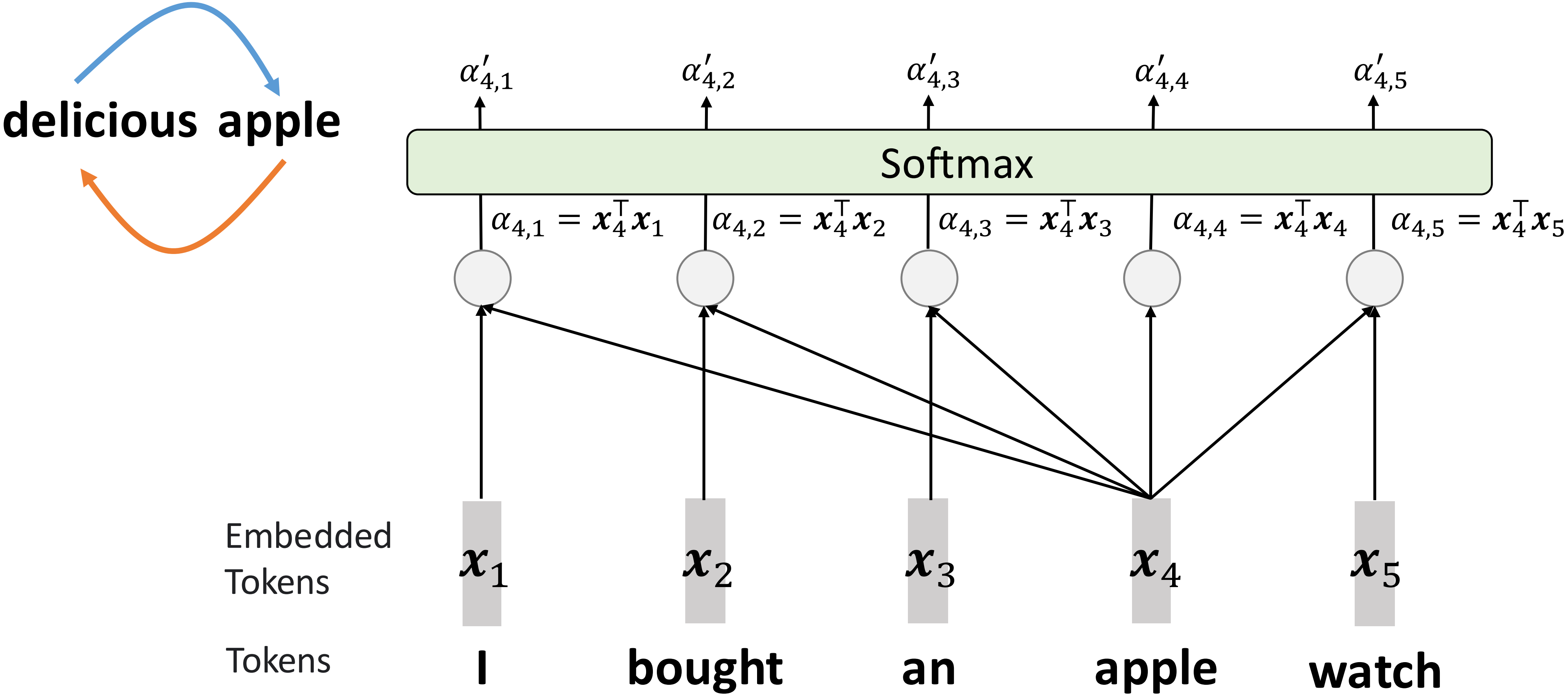
Tokens **I** **bought** **an** **apple** **watch**

Updated feature $\boldsymbol{x}'_4 = \alpha'_{4,1}\boldsymbol{x}_1 + \alpha'_{4,2}\boldsymbol{x}_2 + \alpha'_{4,3}\boldsymbol{x}_3 + \alpha'_{4,4}\boldsymbol{x}_4 + \alpha'_{4,5}\boldsymbol{x}_5$



watch
apple
I
an
bought
$\alpha'_{4,1}$
$\alpha'_{4,2}$
$\alpha'_{4,3}$
$\alpha'_{4,4}$
$\alpha'_{4,5}$
Softmax
$\alpha_{4,1} = \boldsymbol{x}_4^\top \boldsymbol{x}_1$
$\alpha_{4,2} = \boldsymbol{x}_4^\top \boldsymbol{x}_2$
$\alpha_{4,3} = \boldsymbol{x}_4^\top \boldsymbol{x}_3$
$\alpha_{4,4} = \boldsymbol{x}_4^\top \boldsymbol{x}_4$
$\alpha_{4,5} = \boldsymbol{x}_4^\top \boldsymbol{x}_5$
Embedded Tokens
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_3$
$\boldsymbol{x}_4$
$\boldsymbol{x}_5$
Tokens
I
bought
an
apple
watch

Updated feature $\boldsymbol{x}'_4 = \alpha'_{4,1}\boldsymbol{x}_1 + \alpha'_{4,2}\boldsymbol{x}_2 + \alpha'_{4,3}\boldsymbol{x}_3 + \alpha'_{4,4}\boldsymbol{x}_4 + \alpha'_{4,5}\boldsymbol{x}_5$



delicious apple

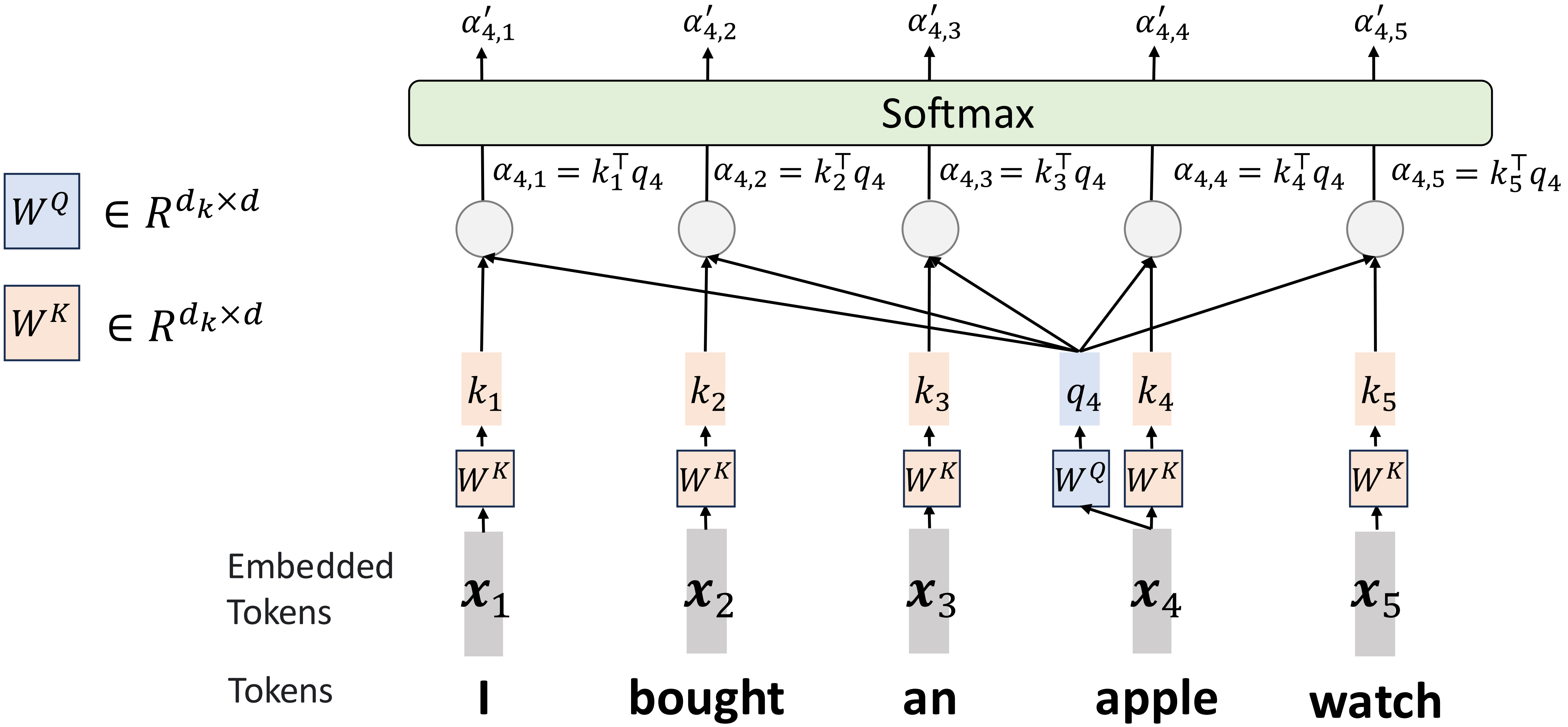
$\alpha'_{4,1}$ $\alpha'_{4,2}$ $\alpha'_{4,3}$ $\alpha'_{4,4}$ $\alpha'_{4,5}$

Softmax

$\alpha_{4,1} = \boldsymbol{x}_4^\top \boldsymbol{x}_1$ $\alpha_{4,2} = \boldsymbol{x}_4^\top \boldsymbol{x}_2$ $\alpha_{4,3} = \boldsymbol{x}_4^\top \boldsymbol{x}_3$ $\alpha_{4,4} = \boldsymbol{x}_4^\top \boldsymbol{x}_4$ $\alpha_{4,5} = \boldsymbol{x}_4^\top \boldsymbol{x}_5$

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens **I** **bought** **an** **apple** **watch**



$W^Q \in R^{d_k \times d}$

$W^K \in R^{d_k \times d}$

$\alpha'_{4,1}$ $\alpha'_{4,2}$ $\alpha'_{4,3}$ $\alpha'_{4,4}$ $\alpha'_{4,5}$

Softmax

$\alpha_{4,1} = k_1^\top q_4$ $\alpha_{4,2} = k_2^\top q_4$ $\alpha_{4,3} = k_3^\top q_4$ $\alpha_{4,4} = k_4^\top q_4$ $\alpha_{4,5} = k_5^\top q_4$

k_1 k_2 k_3 q_4 k_4 k_5

W^K W^K W^K W^Q W^K W^K

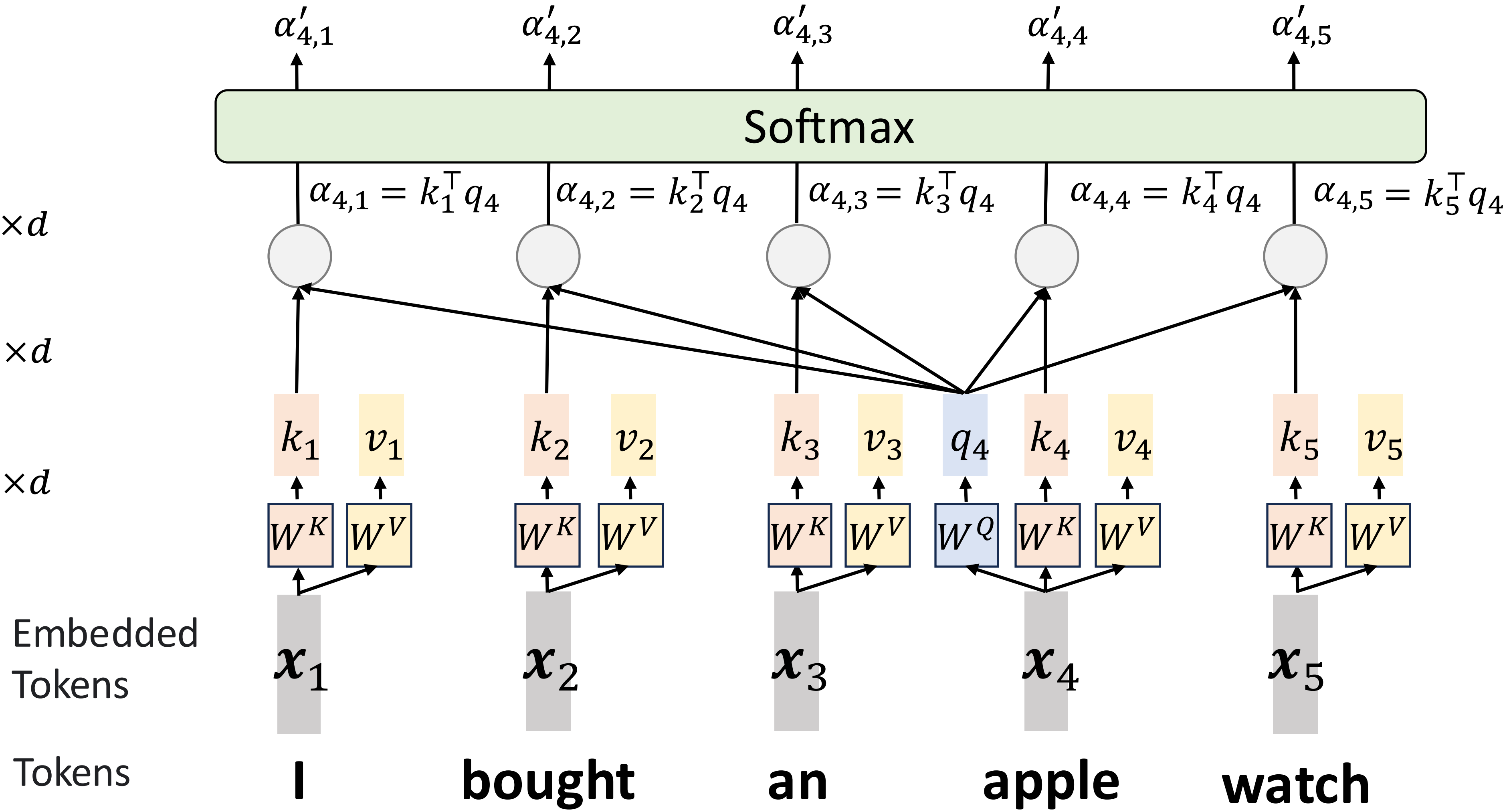
Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens **I** **bought** **an** **apple** **watch**

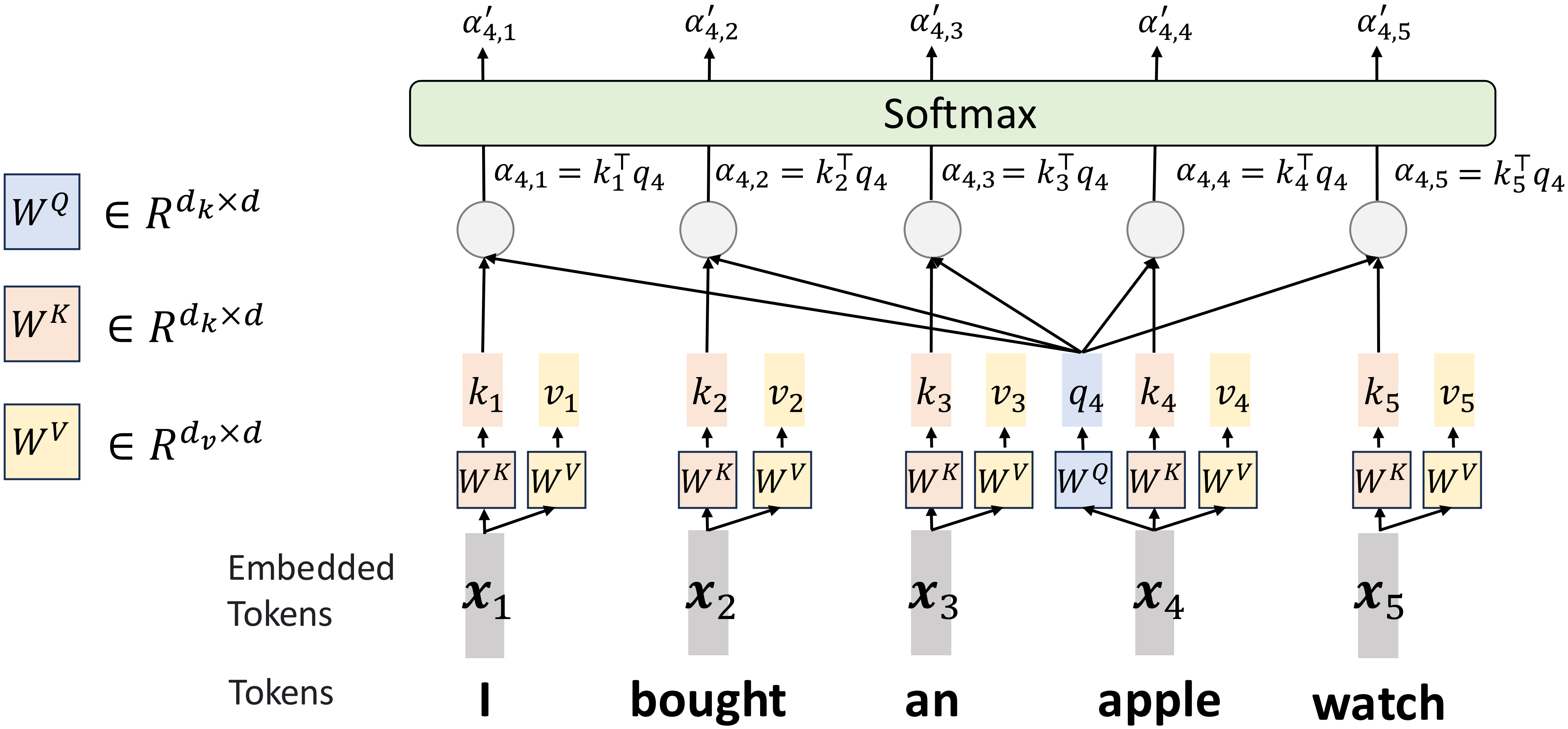
$W^Q \in R^{d_k \times d}$

$W^K \in R^{d_k \times d}$

$W^V \in R^{d_v \times d}$

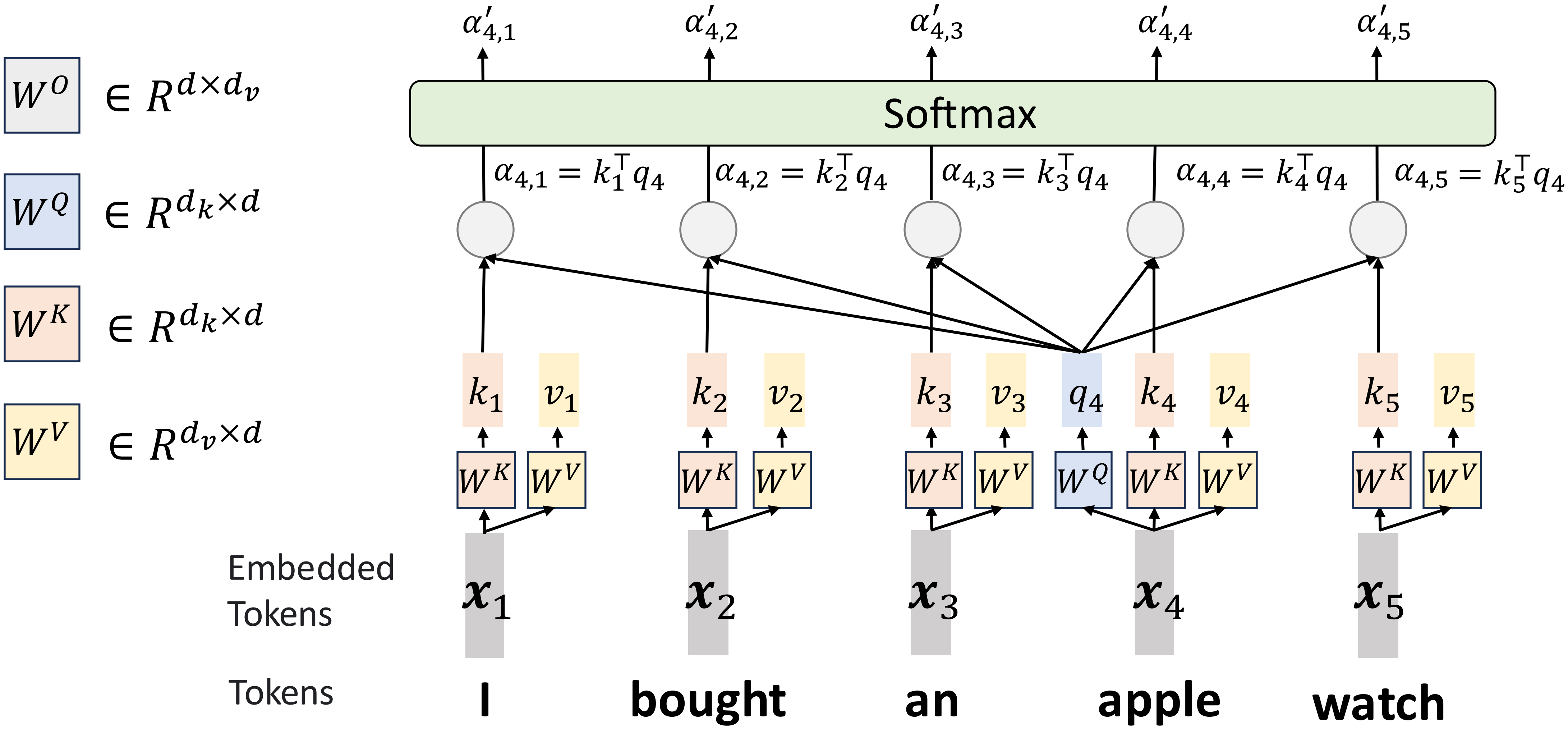


Updated feature $\boldsymbol{x}'_4 = \alpha'_{4,1} v_1 + \alpha'_{4,2} v_2 + \alpha'_{4,3} v_3 + \alpha'_{4,4} v_4 + \alpha'_{4,5} v_5$



Updated feature $$\boldsymbol{x}'_4 = \boxed{W^O}\left(\alpha'_{4,1} v_1 + \alpha'_{4,2} v_2 + \alpha'_{4,3} v_3 + \alpha'_{4,4} v_4 + \alpha'_{4,5} v_5 \right)$$

$$= \sum_i \alpha'_{4,1} \boxed{W^O} \boxed{W^V} \boldsymbol{x}_i$$



$W^O \in R^{d \times d_v}$

$W^Q \in R^{d_k \times d}$

$W^K \in R^{d_k \times d}$

$W^V \in R^{d_v \times d}$

Updated feature $\boldsymbol{x}'_4 = \boxed{W^O}\left(\alpha'_{4,1} v_1 + \alpha'_{4,2} v_2 + \alpha'_{4,3} v_3 + \alpha'_{4,4} v_4 + \alpha'_{4,5} v_5 \right)$

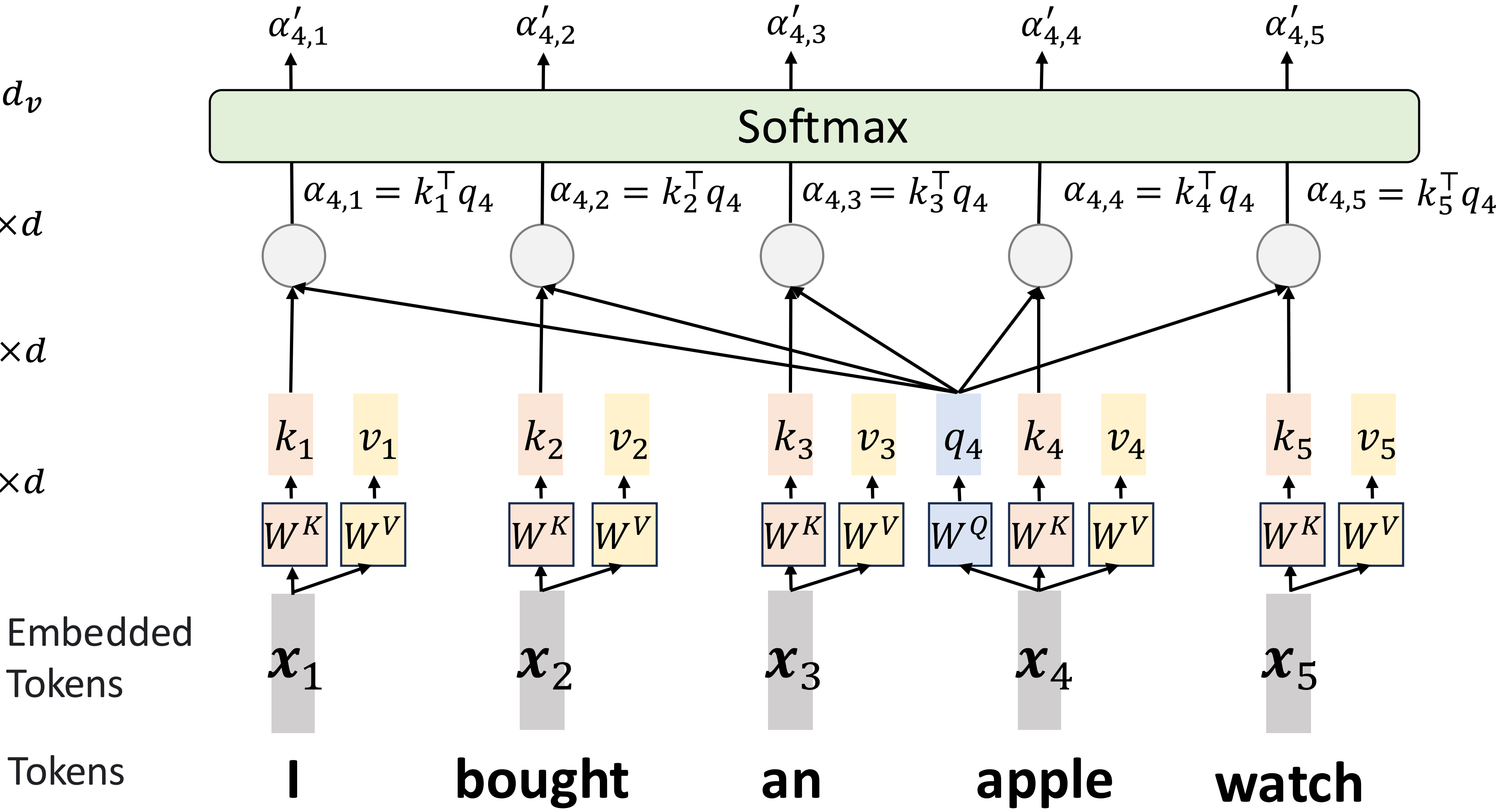
$$= \sum_i \alpha'_{4,1} \left(\boxed{W^O} \boxed{W^V} \right) \boldsymbol{x}_i$$

$\boxed{W^O} \in R^{d \times d_v}$

$\boxed{W^Q} \in R^{d_k \times d}$

$\boxed{W^K} \in R^{d_k \times d}$

$\boxed{W^V} \in R^{d_v \times d}$



$\alpha'_{4,1}$ $\alpha'_{4,2}$ $\alpha'_{4,3}$ $\alpha'_{4,4}$ $\alpha'_{4,5}$

Softmax

$\alpha_{4,1} = k_1^\top q_4$ $\alpha_{4,2} = k_2^\top q_4$ $\alpha_{4,3} = k_3^\top q_4$ $\alpha_{4,4} = k_4^\top q_4$ $\alpha_{4,5} = k_5^\top q_4$

k_1 v_1 k_2 v_2 k_3 v_3 q_4 k_4 v_4 k_5 v_5

W^K W^V W^K W^V W^K W^V W^Q W^K W^V W^K W^V

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

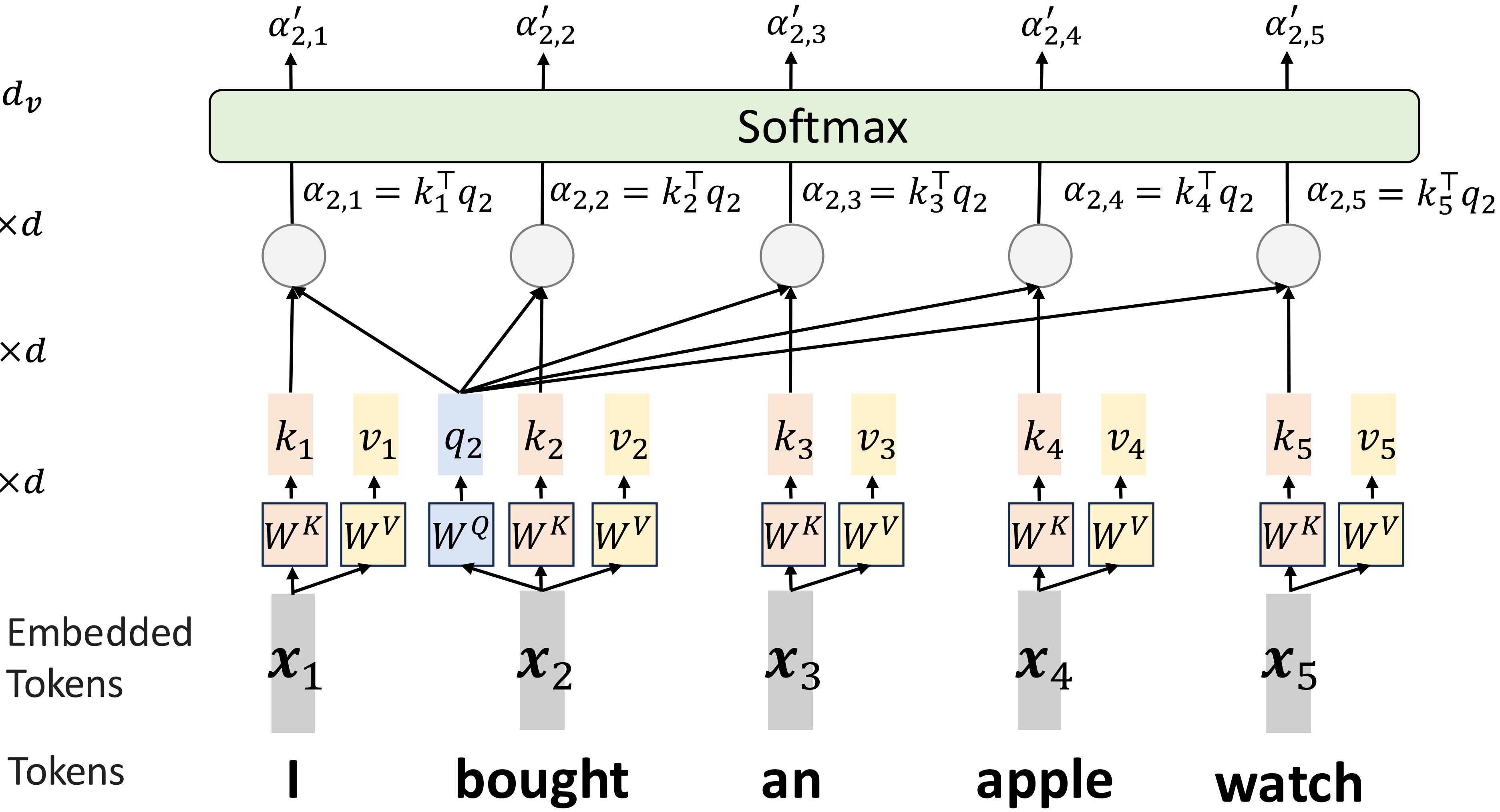
Tokens **I** **bought** **an** **apple** **watch**

$W^O \in R^{d \times d_v}$

$W^Q \in R^{d_k \times d}$

$W^K \in R^{d_k \times d}$

$W^V \in R^{d_v \times d}$



$\alpha'_{2,1}$ $\alpha'_{2,2}$ $\alpha'_{2,3}$ $\alpha'_{2,4}$ $\alpha'_{2,5}$

Softmax

$\alpha_{2,1} = k_1^\top q_2$ $\alpha_{2,2} = k_2^\top q_2$ $\alpha_{2,3} = k_3^\top q_2$ $\alpha_{2,4} = k_4^\top q_2$ $\alpha_{2,5} = k_5^\top q_2$

k_1 v_1 q_2 k_2 v_2 k_3 v_3 k_4 v_4 k_5 v_5

W^K W^V W^Q W^K W^V W^K W^V W^K W^V W^K W^V

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

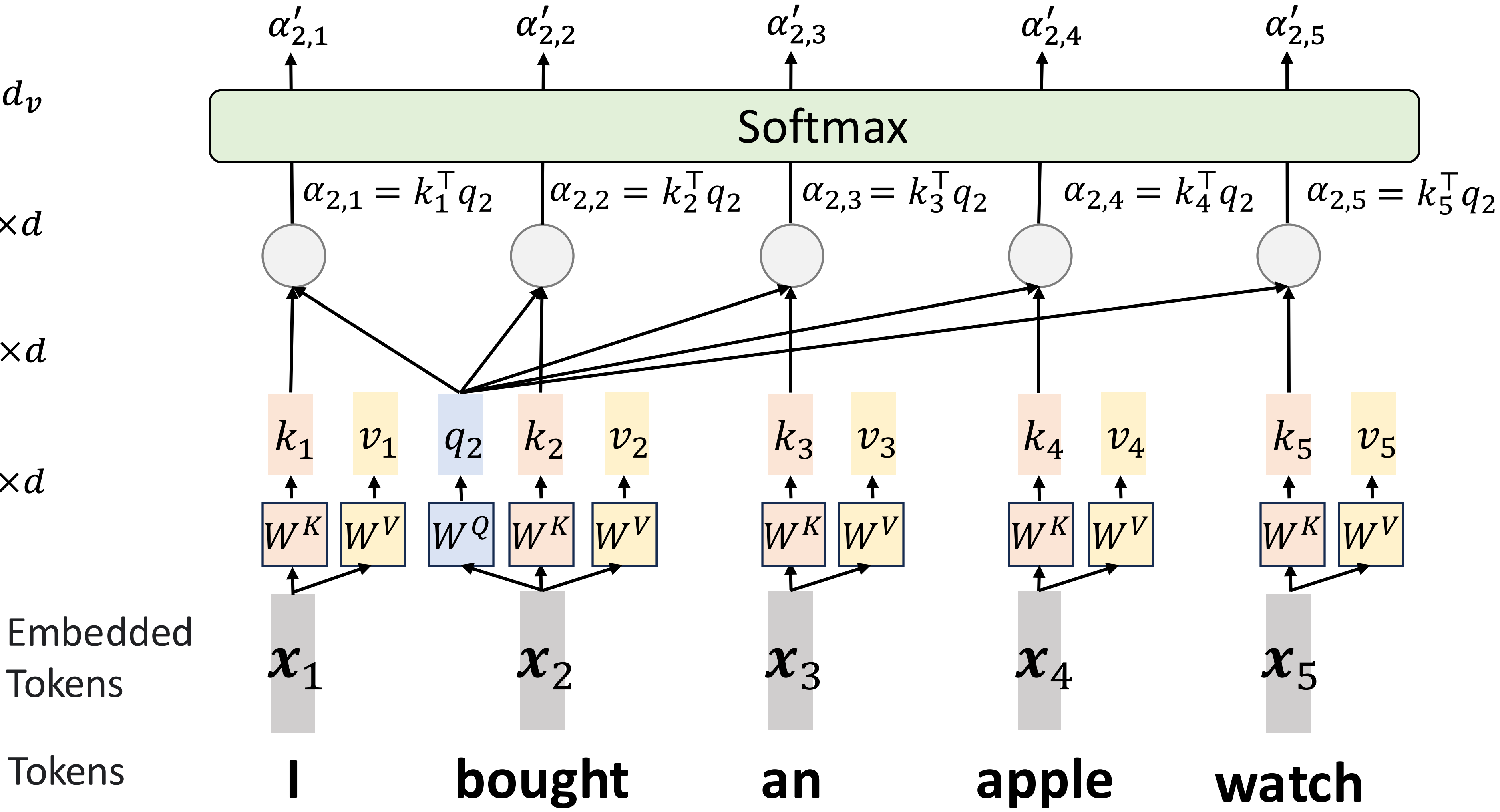
Tokens **I** **bought** **an** **apple** **watch**

$W^O \in R^{d \times d_v}$

$W^Q \in R^{d_k \times d}$

$W^K \in R^{d_k \times d}$

$W^V \in R^{d_v \times d}$



$\alpha'_{2,1}$ $\alpha'_{2,2}$ $\alpha'_{2,3}$ $\alpha'_{2,4}$ $\alpha'_{2,5}$

Softmax

$\alpha_{2,1} = k_1^\top q_2$ $\alpha_{2,2} = k_2^\top q_2$ $\alpha_{2,3} = k_3^\top q_2$ $\alpha_{2,4} = k_4^\top q_2$ $\alpha_{2,5} = k_5^\top q_2$

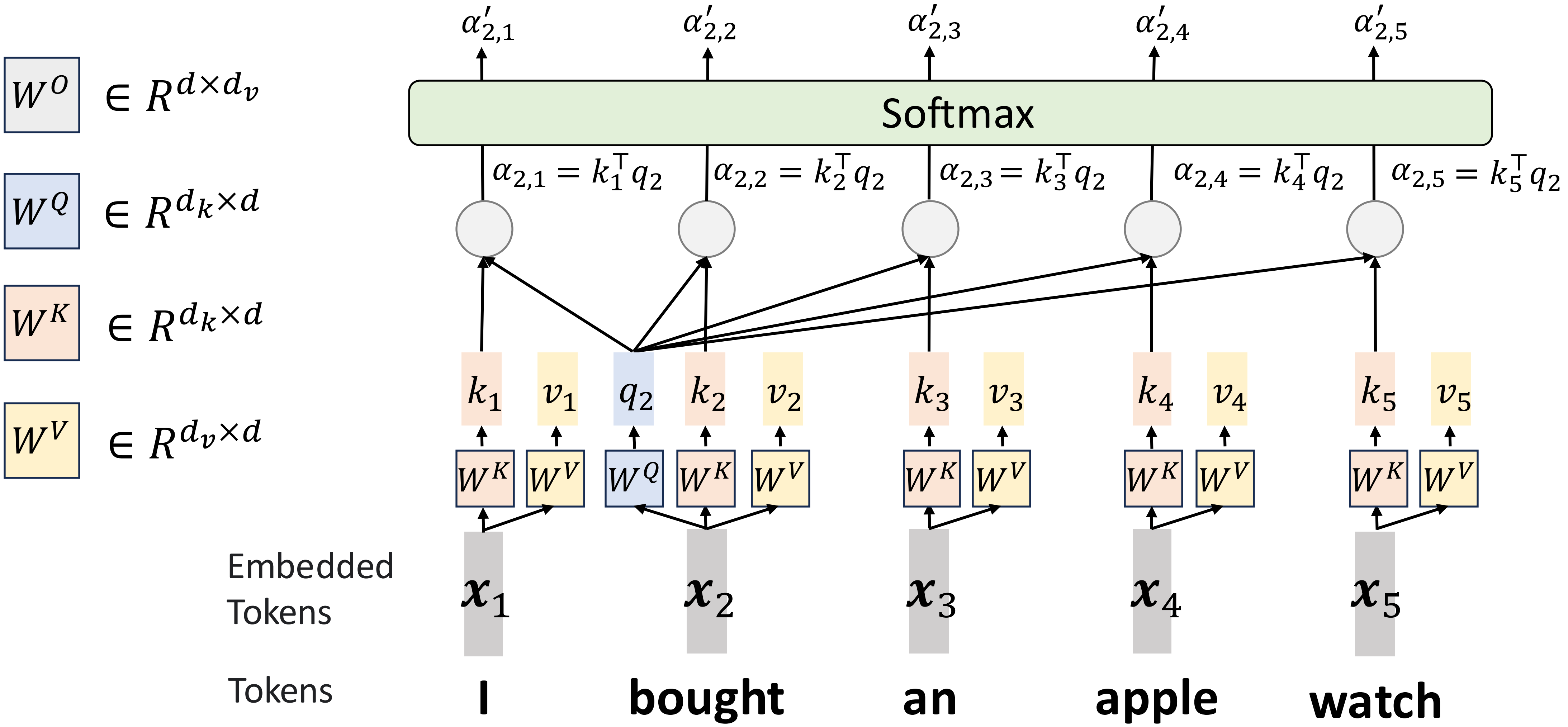
k_1 v_1 q_2 k_2 v_2 k_3 v_3 k_4 v_4 k_5 v_5

W^K W^V W^Q W^K W^V W^K W^V W^K W^V W^K W^V

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens **I** **bought** **an** **apple** **watch**

Updated feature $\boldsymbol{x}'_2 = W^O(\alpha'_{2,1} v_1 + \alpha'_{2,2} v_2 + \alpha'_{2,3} v_3 + \alpha'_{2,4} v_4 + \alpha'_{2,5} v_5)$



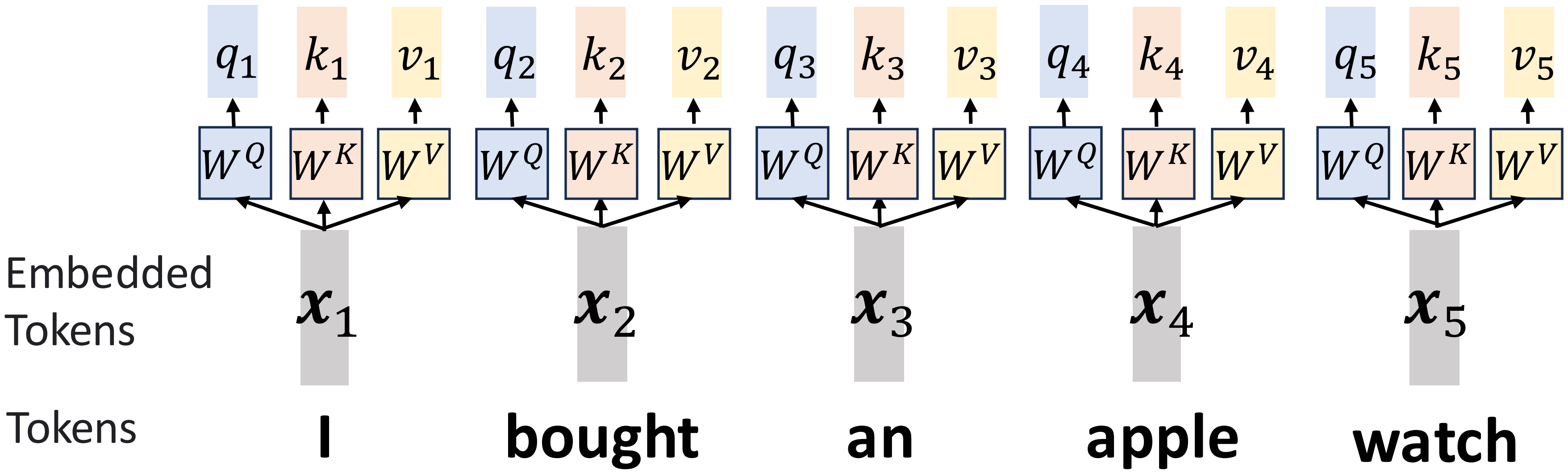
$$\alpha_{1,1} = k_1^\top q_1$$

$$\alpha_{1,2} = k_2^\top q_1$$

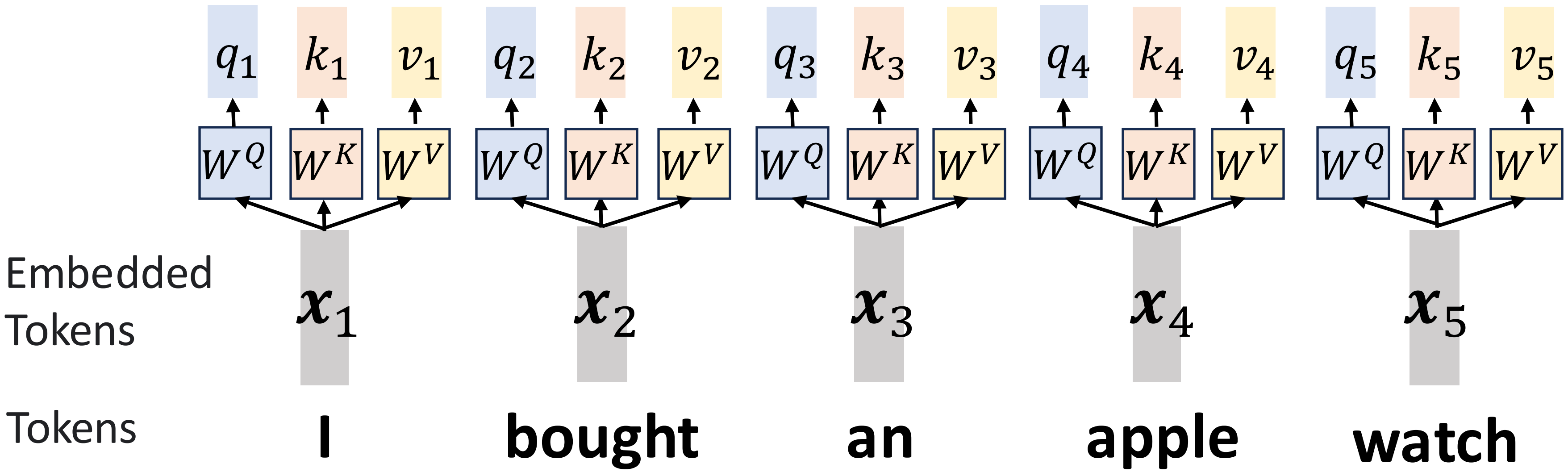
$$\alpha_{1,3} = k_3^\top q_1$$

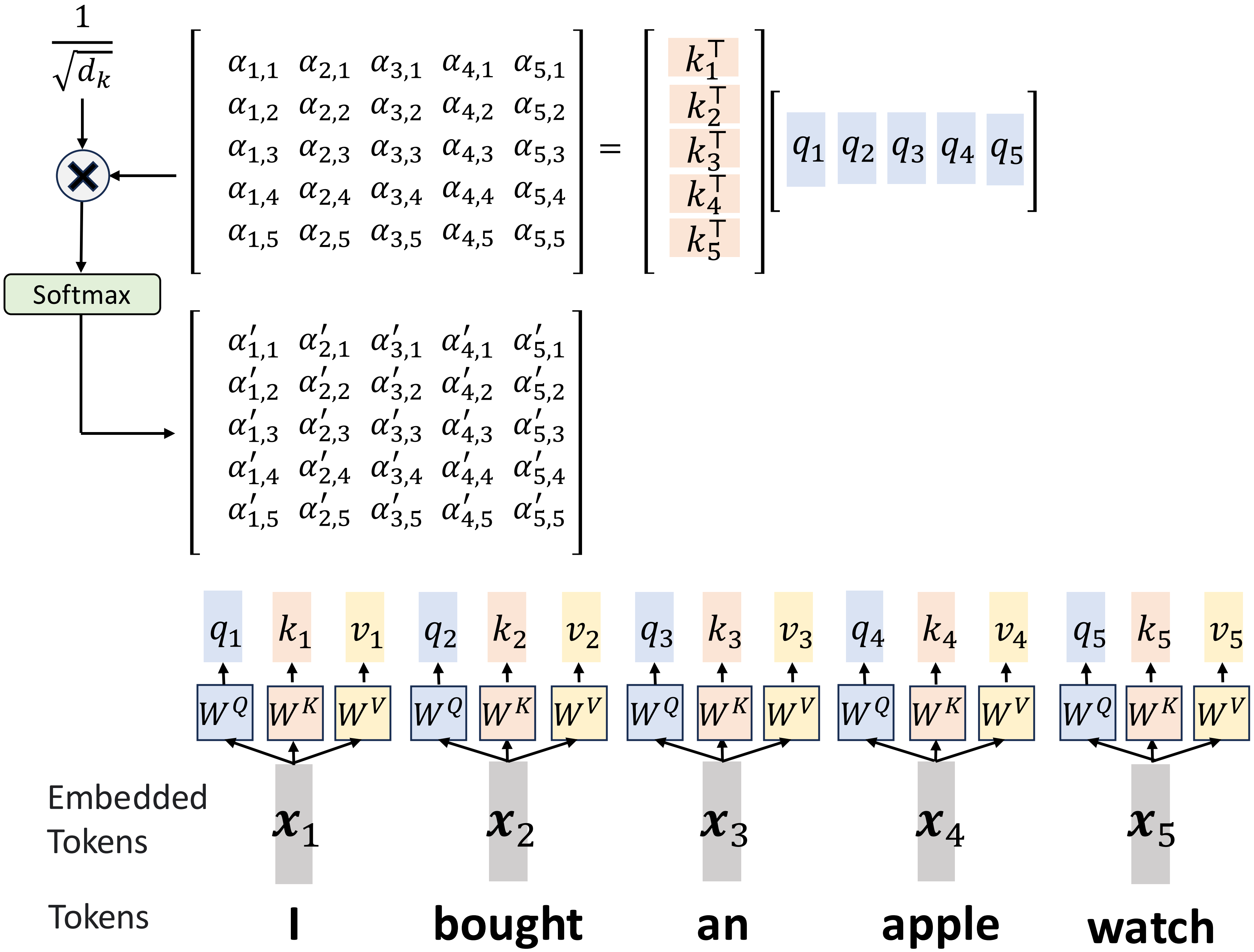
$$\alpha_{1,4} = k_4^\top q_1$$

$$\alpha_{1,5} = k_5^\top q_1$$


q_1 k_1 v_1 q_2 k_2 v_2 q_3 k_3 v_3 q_4 k_4 v_4 q_5 k_5 v_5
W^Q W^K W^V
Embedded Tokens
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$
Tokens
I
bought
an
apple
watch

$$\begin{matrix} \alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1} & \alpha_{5,1} \\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2} & \alpha_{5,2} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3} & \alpha_{5,3} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4} & \alpha_{5,4} \\ \alpha_{1,5} & \alpha_{2,5} & \alpha_{3,5} & \alpha_{4,5} & \alpha_{5,5} \end{matrix} = \begin{bmatrix} k_1^\top \\ k_2^\top \\ k_3^\top \\ k_4^\top \\ k_5^\top \end{bmatrix} \begin{matrix} q_1 & q_2 & q_3 & q_4 & q_5 \end{matrix}$$


q_1 k_1 v_1 q_2 k_2 v_2 q_3 k_3 v_3 q_4 k_4 v_4 q_5 k_5 v_5
W^Q W^K W^V
Embedded Tokens
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$
Tokens
I
bought
an
apple
watch

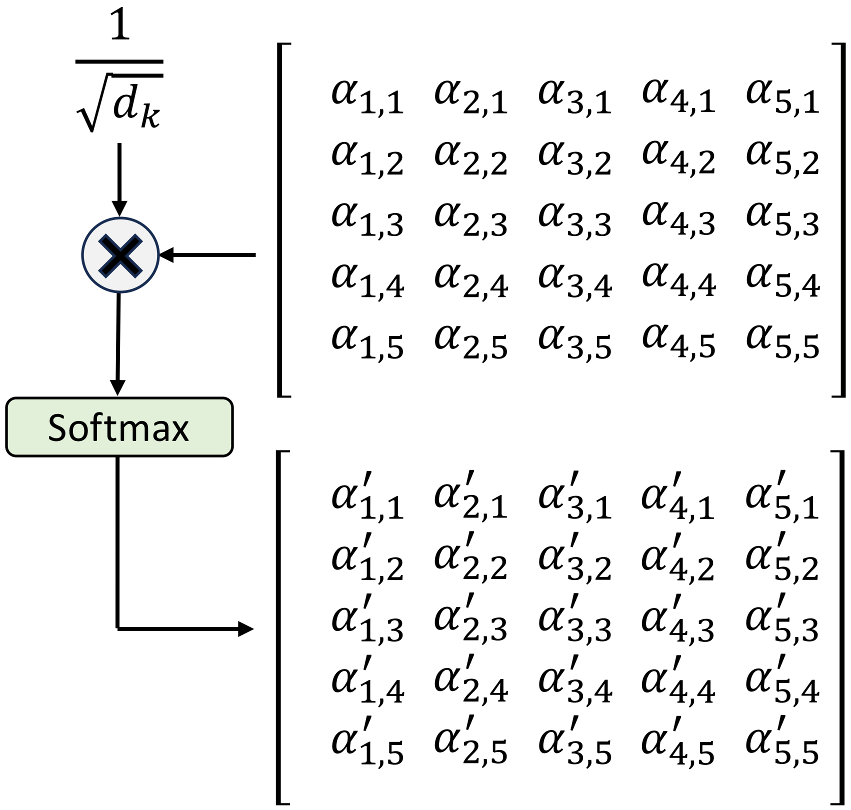


$$\frac{1}{\sqrt{d_k}}$$

$$\begin{bmatrix} \alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1} & \alpha_{5,1} \\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2} & \alpha_{5,2} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3} & \alpha_{5,3} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4} & \alpha_{5,4} \\ \alpha_{1,5} & \alpha_{2,5} & \alpha_{3,5} & \alpha_{4,5} & \alpha_{5,5} \end{bmatrix} = \begin{bmatrix} k_1^\top \\ k_2^\top \\ k_3^\top \\ k_4^\top \\ k_5^\top \end{bmatrix} \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \end{bmatrix}$$

$$\begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} & \alpha'_{5,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} & \alpha'_{5,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} & \alpha'_{5,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} & \alpha'_{5,4} \\ \alpha'_{1,5} & \alpha'_{2,5} & \alpha'_{3,5} & \alpha'_{4,5} & \alpha'_{5,5} \end{bmatrix}$$

Softmax
q_1 k_1 v_1 q_2 k_2 v_2 q_3 k_3 v_3 q_4 k_4 v_4 q_5 k_5 v_5
W^Q W^K W^V
Embedded Tokens
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$
Tokens
I
bought
an
apple
watch



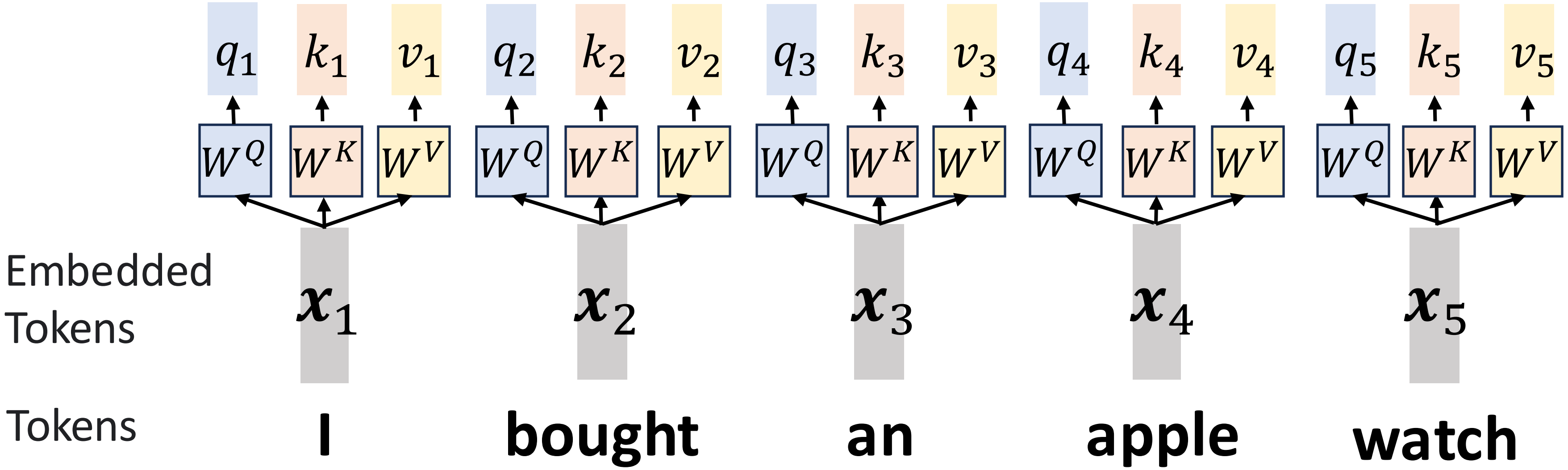
$$\begin{bmatrix} \alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1} & \alpha_{5,1} \\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2} & \alpha_{5,2} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3} & \alpha_{5,3} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4} & \alpha_{5,4} \\ \alpha_{1,5} & \alpha_{2,5} & \alpha_{3,5} & \alpha_{4,5} & \alpha_{5,5} \end{bmatrix} = \begin{bmatrix} k_1^\top \\ k_2^\top \\ k_3^\top \\ k_4^\top \\ k_5^\top \end{bmatrix} \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \end{bmatrix}$$

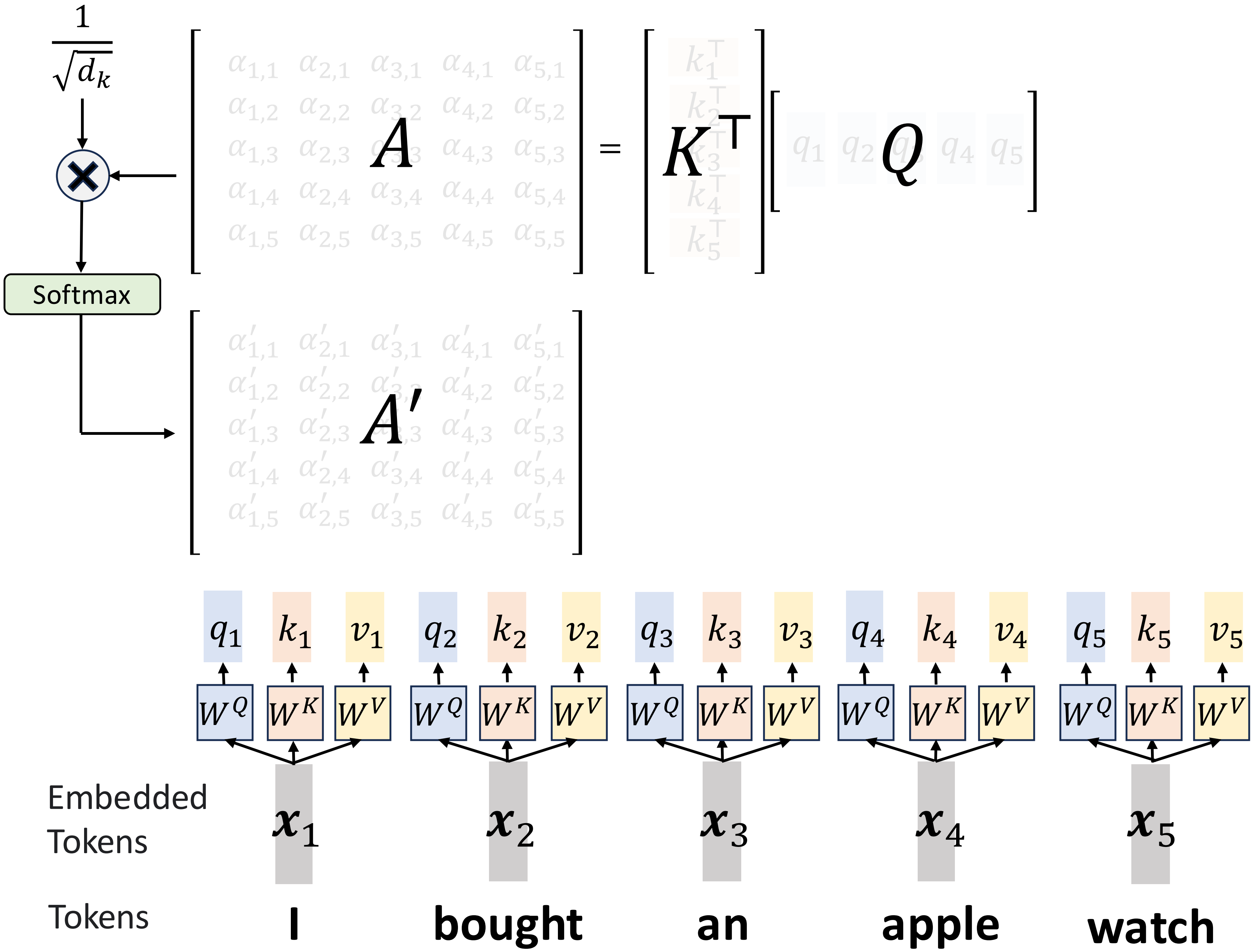
$$\begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} & \alpha'_{5,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} & \alpha'_{5,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} & \alpha'_{5,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} & \alpha'_{5,4} \\ \alpha'_{1,5} & \alpha'_{2,5} & \alpha'_{3,5} & \alpha'_{4,5} & \alpha'_{5,5} \end{bmatrix}$$

$$k = [k^1, k^2, \cdots, k^{d_k}]^\top \qquad E[k^i] = E[q^i] = 0$$

$$q = [q^1, q^2, \cdots, q^{d_k}]^\top \qquad \mathrm{Var}[k^i] = \mathrm{Var}[q^i] = 1$$

$$k^\top q = \sum_{i=1}^{d_k} k^i q^i \qquad \mathrm{Var}[k^\top q] = d_k$$

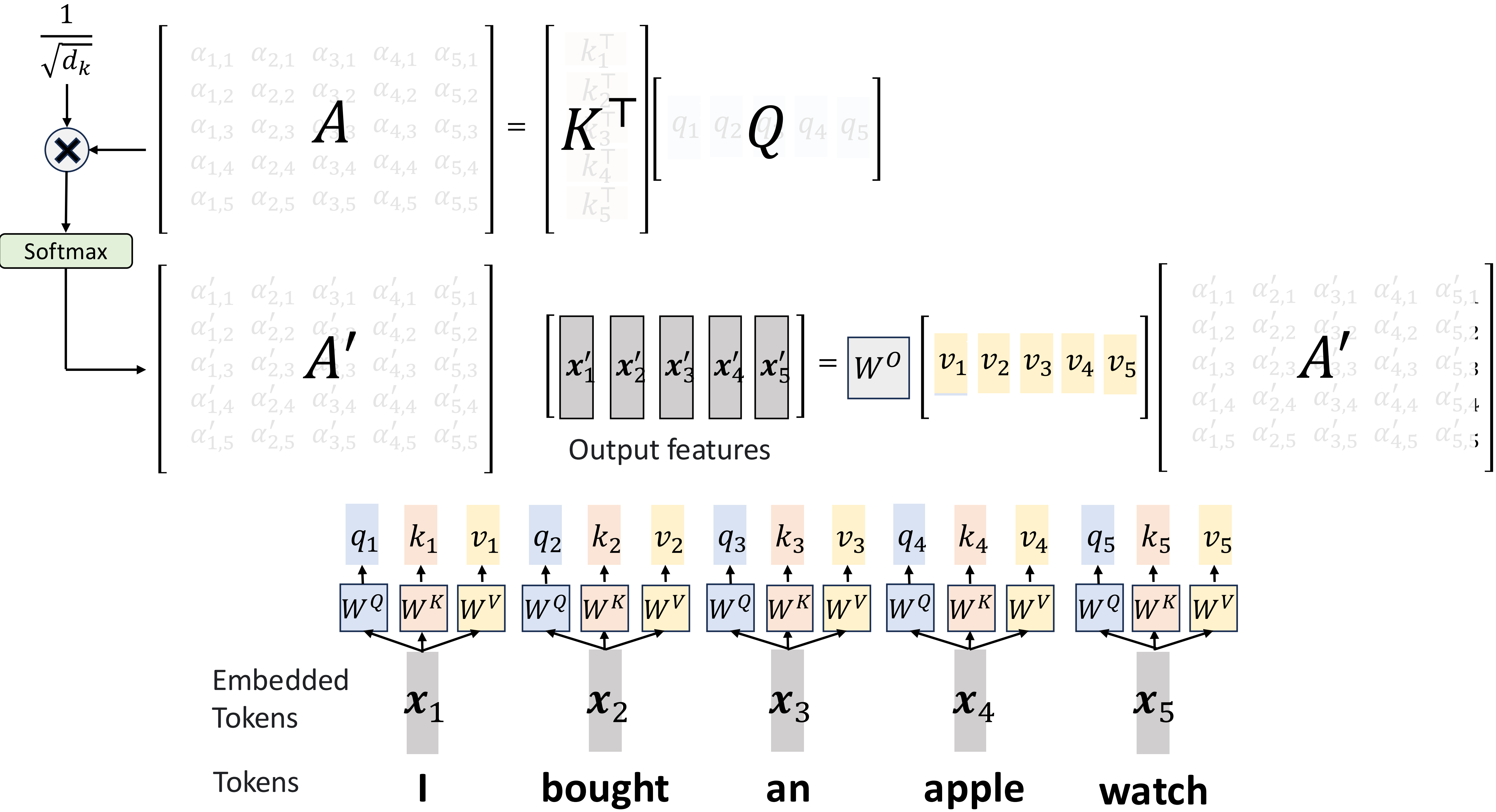

q_1 k_1 v_1 q_2 k_2 v_2 q_3 k_3 v_3 q_4 k_4 v_4 q_5 k_5 v_5
W^Q W^K W^V
Embedded Tokens
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$
Tokens
I
bought
an
apple
watch



$\frac{1}{\sqrt{d_k}}$
Softmax
q_1 k_1 v_1 q_2 k_2 v_2 q_3 k_3 v_3 q_4 k_4 v_4 q_5 k_5 v_5
W^Q W^K W^V
Embedded Tokens
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$
Tokens
I
bought
an
apple
watch

$$A = \begin{bmatrix} \alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1} & \alpha_{5,1} \\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2} & \alpha_{5,2} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3} & \alpha_{5,3} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4} & \alpha_{5,4} \\ \alpha_{1,5} & \alpha_{2,5} & \alpha_{3,5} & \alpha_{4,5} & \alpha_{5,5} \end{bmatrix} = K^\top Q = \begin{bmatrix} k_1^\top \\ k_2^\top \\ k_3^\top \\ k_4^\top \\ k_5^\top \end{bmatrix} \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \end{bmatrix}$$

$$A' = \begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} & \alpha'_{5,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} & \alpha'_{5,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} & \alpha'_{5,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} & \alpha'_{5,4} \\ \alpha'_{1,5} & \alpha'_{2,5} & \alpha'_{3,5} & \alpha'_{4,5} & \alpha'_{5,5} \end{bmatrix}$$



$$\frac{1}{\sqrt{d_k}}$$

Softmax

$$A = \begin{bmatrix} \alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1} & \alpha_{5,1} \\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2} & \alpha_{5,2} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3} & \alpha_{5,3} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4} & \alpha_{5,4} \\ \alpha_{1,5} & \alpha_{2,5} & \alpha_{3,5} & \alpha_{4,5} & \alpha_{5,5} \end{bmatrix} = K^\top Q = \begin{bmatrix} k_1^\top \\ k_2^\top \\ k_3^\top \\ k_4^\top \\ k_5^\top \end{bmatrix} \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \end{bmatrix}$$

$$A' = \begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} & \alpha'_{5,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} & \alpha'_{5,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} & \alpha'_{5,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} & \alpha'_{5,4} \\ \alpha'_{1,5} & \alpha'_{2,5} & \alpha'_{3,5} & \alpha'_{4,5} & \alpha'_{5,5} \end{bmatrix}$$

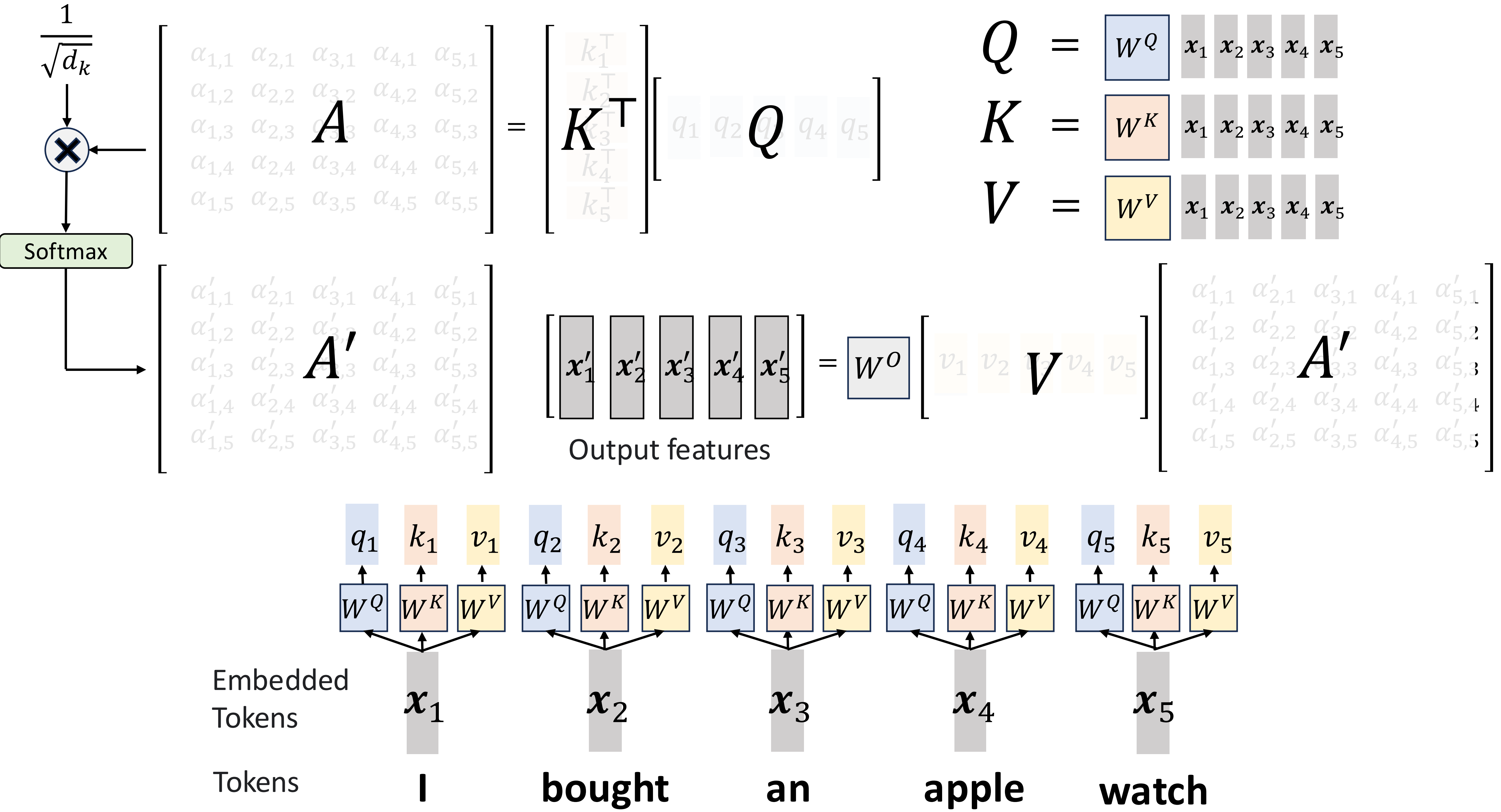
$$\begin{bmatrix} x'_1 & x'_2 & x'_3 & x'_4 & x'_5 \end{bmatrix} = W^O \begin{bmatrix} v_1 & v_2 & v_3 & v_4 & v_5 \end{bmatrix} A'$$

Output features

q_1 k_1 v_1 q_2 k_2 v_2 q_3 k_3 v_3 q_4 k_4 v_4 q_5 k_5 v_5
W^Q W^K W^V

Embedded Tokens $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens **I** **bought** **an** **apple** **watch**



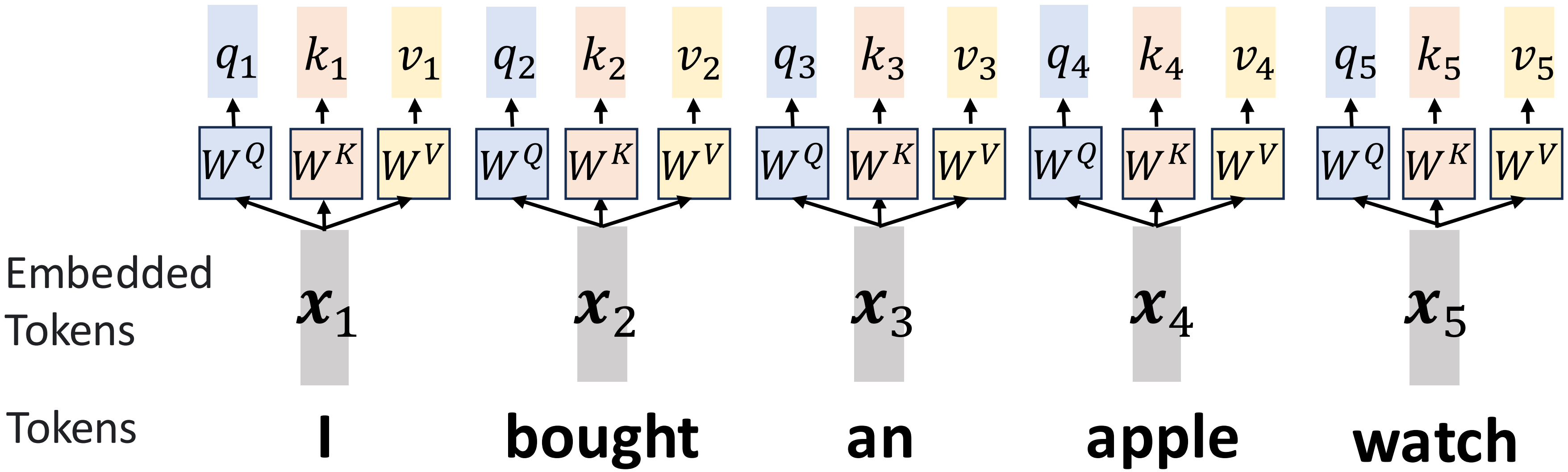
Single-head attention

$$\mathrm{Attention}(Q,K,V) = V\,\mathrm{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$

$$Q = \boxed{W^Q}\; \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4 & \boldsymbol{x}_5 \end{bmatrix}$$

$$K = \boxed{W^K}\; \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4 & \boldsymbol{x}_5 \end{bmatrix}$$

$$V = \boxed{W^V}\; \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4 & \boldsymbol{x}_5 \end{bmatrix}$$


q_1 k_1 v_1 q_2 k_2 v_2 q_3 k_3 v_3 q_4 k_4 v_4 q_5 k_5 v_5
W^Q W^K W^V
Embedded Tokens
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$
Tokens
I
bought
an
apple
watch

Single-head attention

$$\text{Attention}(Q, K, V) = V\ \text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$

$$Q = W^Q\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

$$K = W^K\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

$$V = W^V\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

W_0^Q W_0^K W_0^V W_1^Q W_1^K W_1^V … W_{h-1}^Q W_{h-1}^K W_{h-1}^V

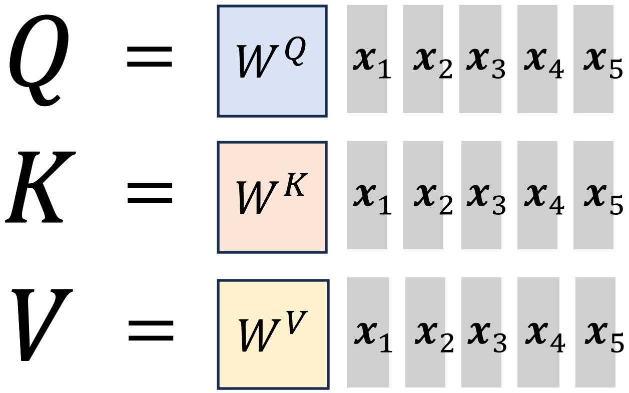
$W_i^Q \in R^{d_k \times d}$

$W_i^K \in R^{d_k \times d}$

$W_i^V \in R^{d_v \times d}$

Single-head attention

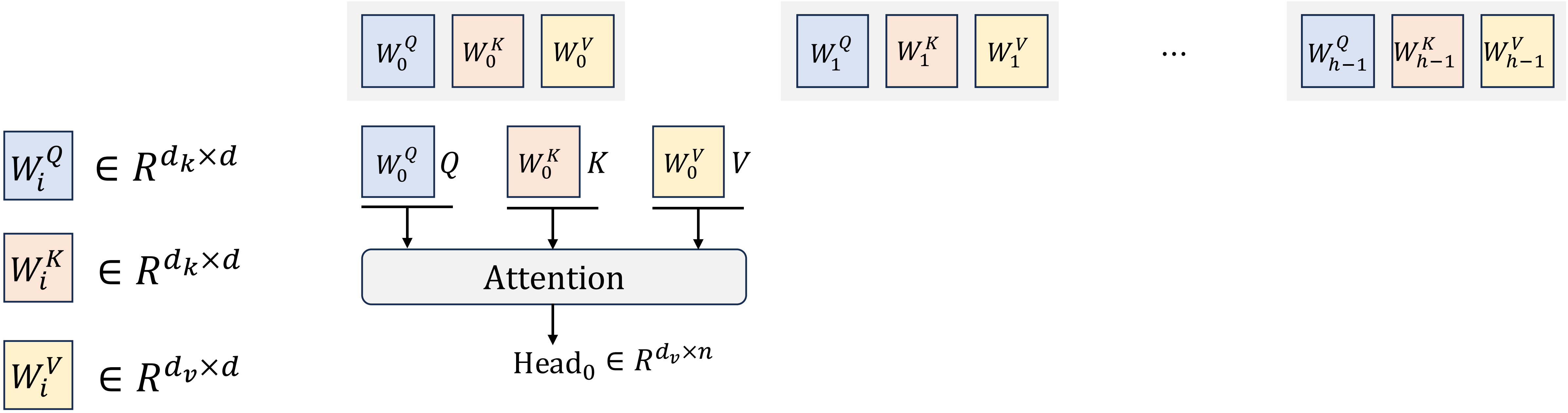
$$\text{Attention}(Q, K, V) = V \,\text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$



$$Q = W^Q \; x_1 \; x_2 \; x_3 \; x_4 \; x_5$$

$$K = W^K \; x_1 \; x_2 \; x_3 \; x_4 \; x_5$$

$$V = W^V \; x_1 \; x_2 \; x_3 \; x_4 \; x_5$$



W_0^Q W_0^K W_0^V

W_1^Q W_1^K W_1^V

…

W_{h-1}^Q W_{h-1}^K W_{h-1}^V

$W_i^Q \in R^{d_k \times d}$

$W_i^K \in R^{d_k \times d}$

$W_i^V \in R^{d_v \times d}$

$W_0^Q Q$ $W_0^K K$ $W_0^V V$

Attention

$\text{Head}_0 \in R^{d_v \times n}$

Single-head attention

$$\text{Attention}(Q, K, V) = V\ \text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$

$$Q = W^Q \quad \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

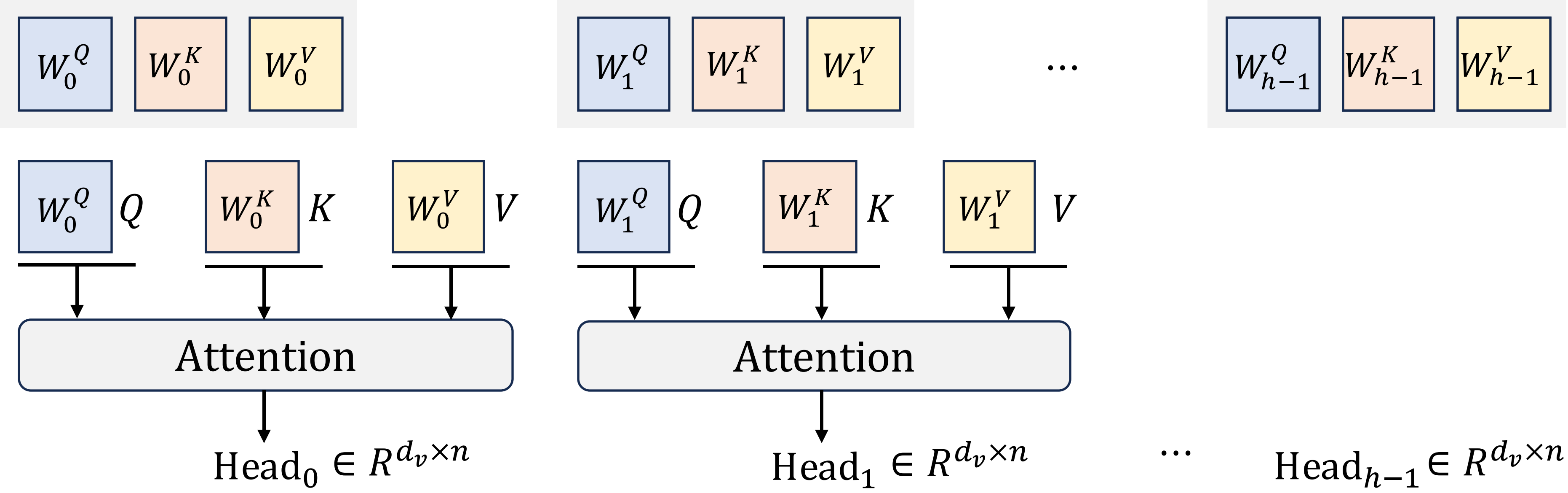
$$K = W^K \quad \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

$$V = W^V \quad \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

$W_i^Q \in R^{d_k \times d}$

$W_i^K \in R^{d_k \times d}$

$W_i^V \in R^{d_v \times d}$



W_0^Q W_0^K W_0^V

W_1^Q W_1^K W_1^V

…

W_{h-1}^Q W_{h-1}^K W_{h-1}^V

$W_0^Q Q$ $W_0^K K$ $W_0^V V$

$W_1^Q Q$ $W_1^K K$ $W_1^V V$

Attention

Attention

$\text{Head}_0 \in R^{d_v \times n}$

$\text{Head}_1 \in R^{d_v \times n}$

…

$\text{Head}_{h-1} \in R^{d_v \times n}$

Single-head attention

$$\text{Attention}(Q,K,V) = V\ \text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$

$$Q = W^Q\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

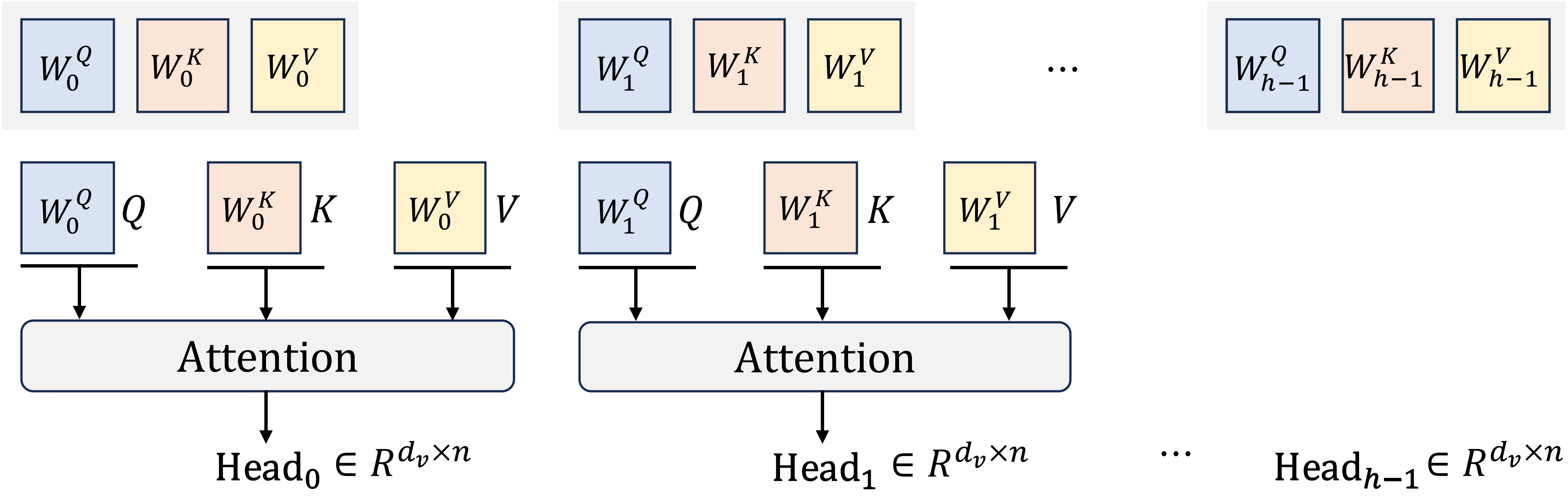
$$K = W^K\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

$$V = W^V\ \boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3\ \boldsymbol{x}_4\ \boldsymbol{x}_5$$

$W_i^Q \in R^{d_k \times d}$

$W_i^K \in R^{d_k \times d}$

$W_i^V \in R^{d_v \times d}$



W_0^Q W_0^K W_0^V

W_1^Q W_1^K W_1^V

…

W_{h-1}^Q W_{h-1}^K W_{h-1}^V

$W_0^Q Q$ $W_0^K K$ $W_0^V V$

$W_1^Q Q$ $W_1^K K$ $W_1^V V$

Attention

Attention

$\text{Head}_0 \in R^{d_v \times n}$

$\text{Head}_1 \in R^{d_v \times n}$

…

$\text{Head}_{h-1} \in R^{d_v \times n}$

Single-head attention

$$\text{Attention}(Q,K,V) = V\,\text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$

$$Q = W^Q \; \boldsymbol{x}_1 \; \boldsymbol{x}_2 \; \boldsymbol{x}_3 \; \boldsymbol{x}_4 \; \boldsymbol{x}_5$$

$$K = W^K \; \boldsymbol{x}_1 \; \boldsymbol{x}_2 \; \boldsymbol{x}_3 \; \boldsymbol{x}_4 \; \boldsymbol{x}_5$$

$$V = W^V \; \boldsymbol{x}_1 \; \boldsymbol{x}_2 \; \boldsymbol{x}_3 \; \boldsymbol{x}_4 \; \boldsymbol{x}_5$$

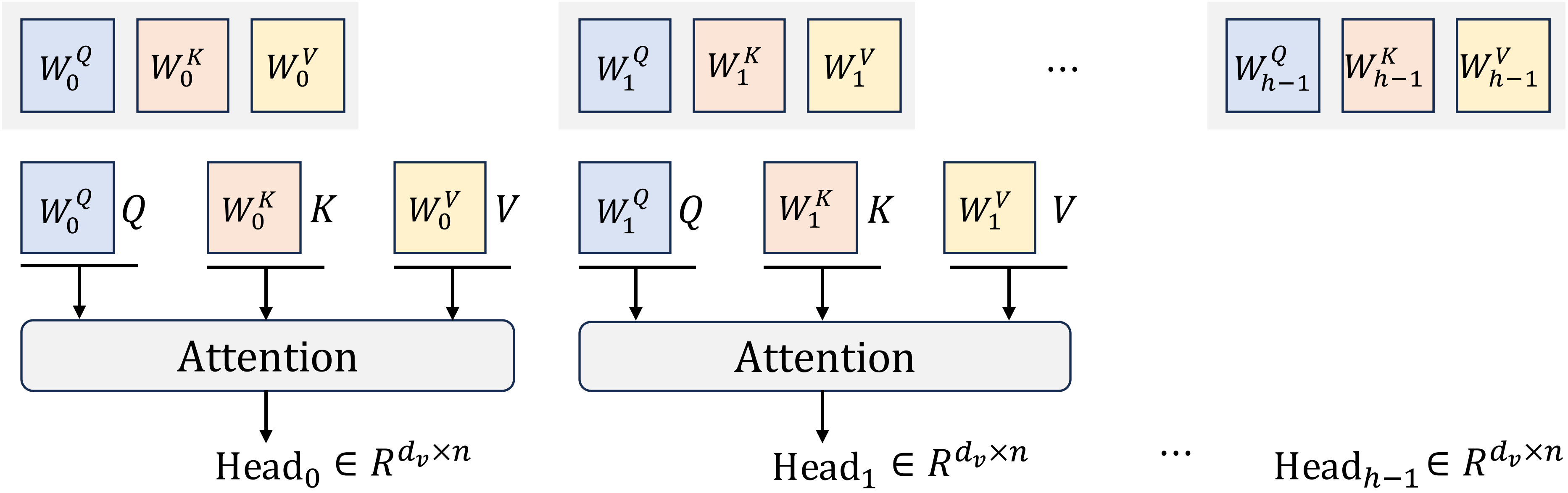
Multi-head attention

$W^O \in R^{d \times hd_v}$

$W_i^Q \in R^{d_k \times d}$

$W_i^K \in R^{d_k \times d}$

$W_i^V \in R^{d_v \times d}$

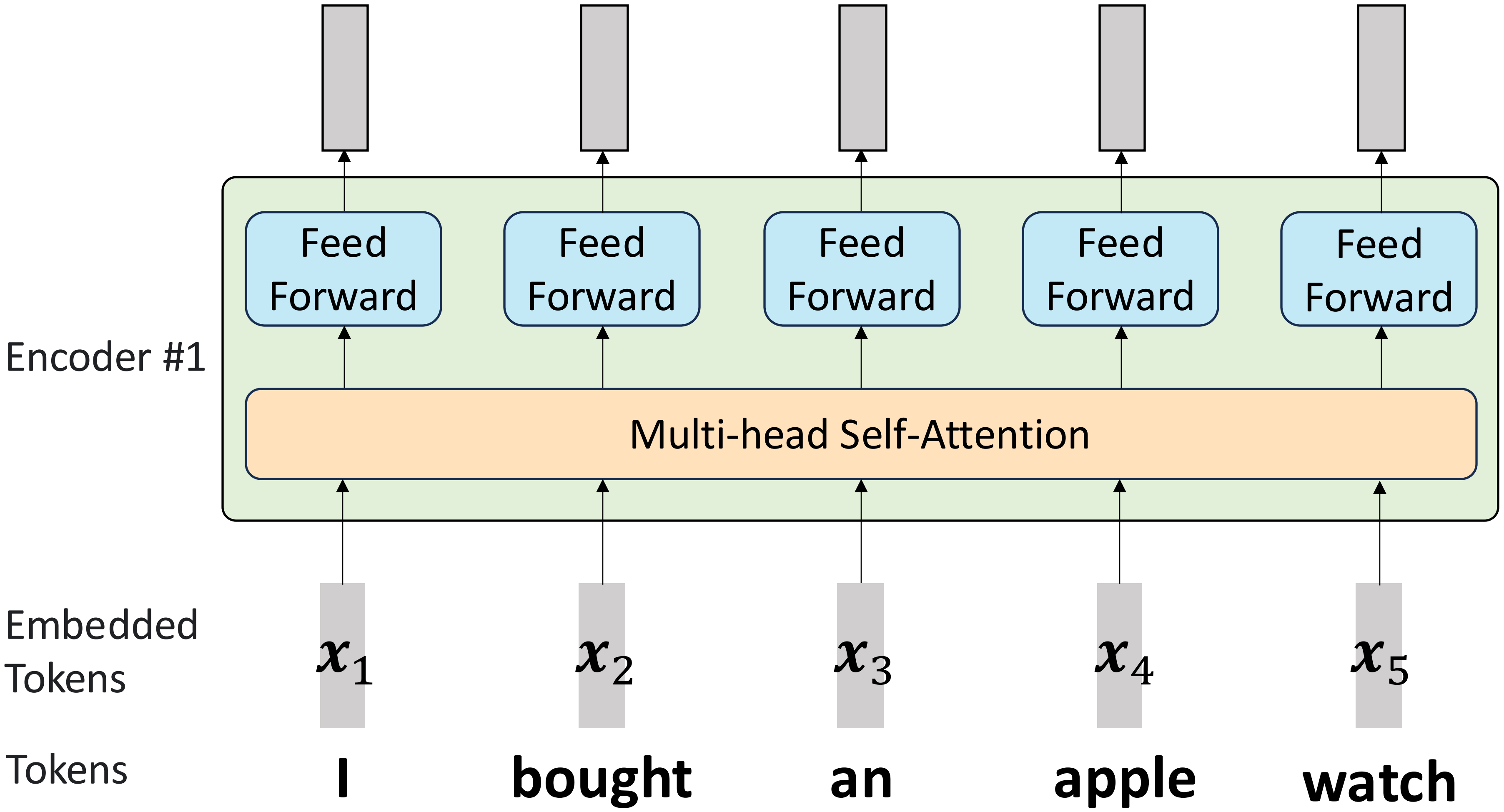


W_0^Q W_0^K W_0^V
W_1^Q W_1^K W_1^V
…
W_{h-1}^Q W_{h-1}^K W_{h-1}^V
$W_0^Q Q$ $W_0^K K$ $W_0^V V$
$W_1^Q Q$ $W_1^K K$ $W_1^V V$
Attention
Attention
$\text{Head}_0 \in R^{d_v \times n}$
$\text{Head}_1 \in R^{d_v \times n}$
…
$\text{Head}_{h-1} \in R^{d_v \times n}$

$$\text{MultiHeadedAttention}(Q,K,V) = W^O \begin{bmatrix} \text{Head}_0 \\ \text{Head}_1 \\ \vdots \\ \text{Head}_{h-1} \end{bmatrix}$$

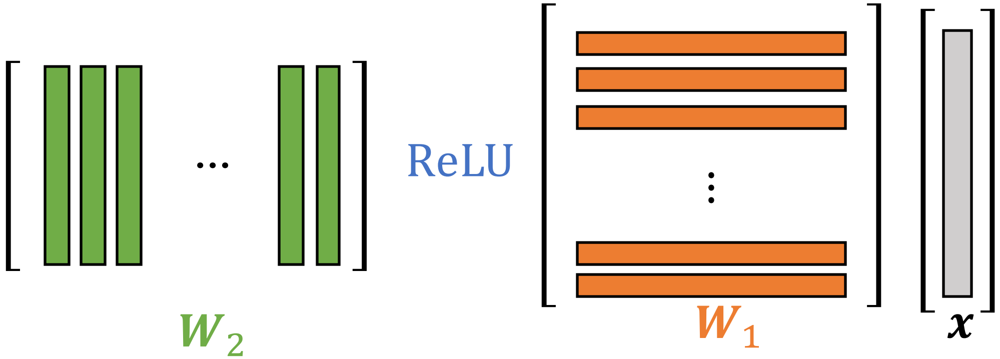
Feed Forward Network (FFN)

$$FFN(\boldsymbol{x}) = \boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2$$

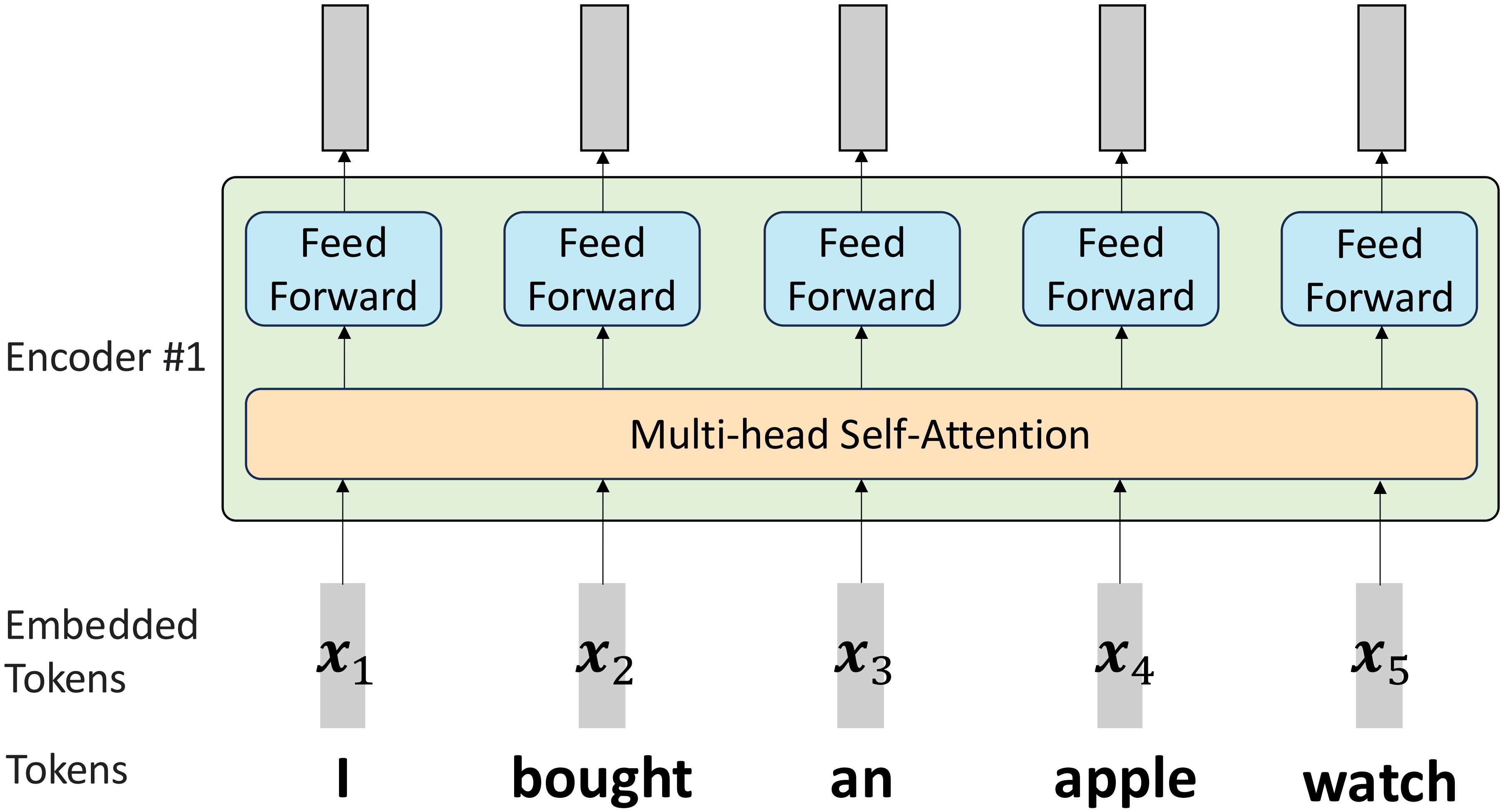


Feed Forward Network (FFN)

$$FFN(\boldsymbol{x}) = \boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2$$



$\boldsymbol{W}_2$ ReLU $\boldsymbol{W}_1$ $\boldsymbol{x}$



Encoder #1

Feed Forward

Feed Forward

Feed Forward

Feed Forward

Feed Forward

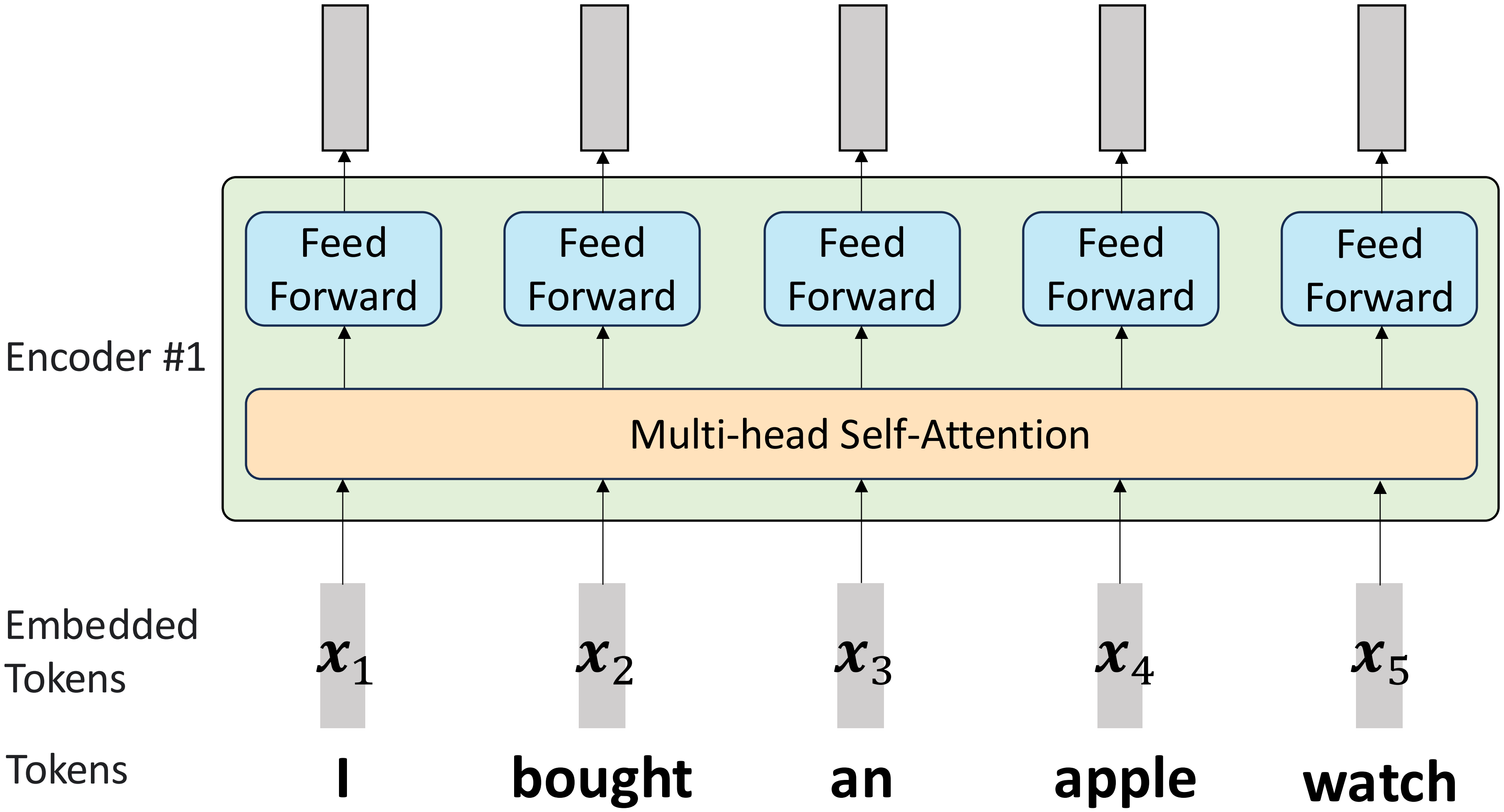
Multi-head Self-Attention

Embedded Tokens

$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens

I **bought** **an** **apple** **watch**



Encoder #1

Feed Forward

Feed Forward

Feed Forward

Feed Forward

Feed Forward

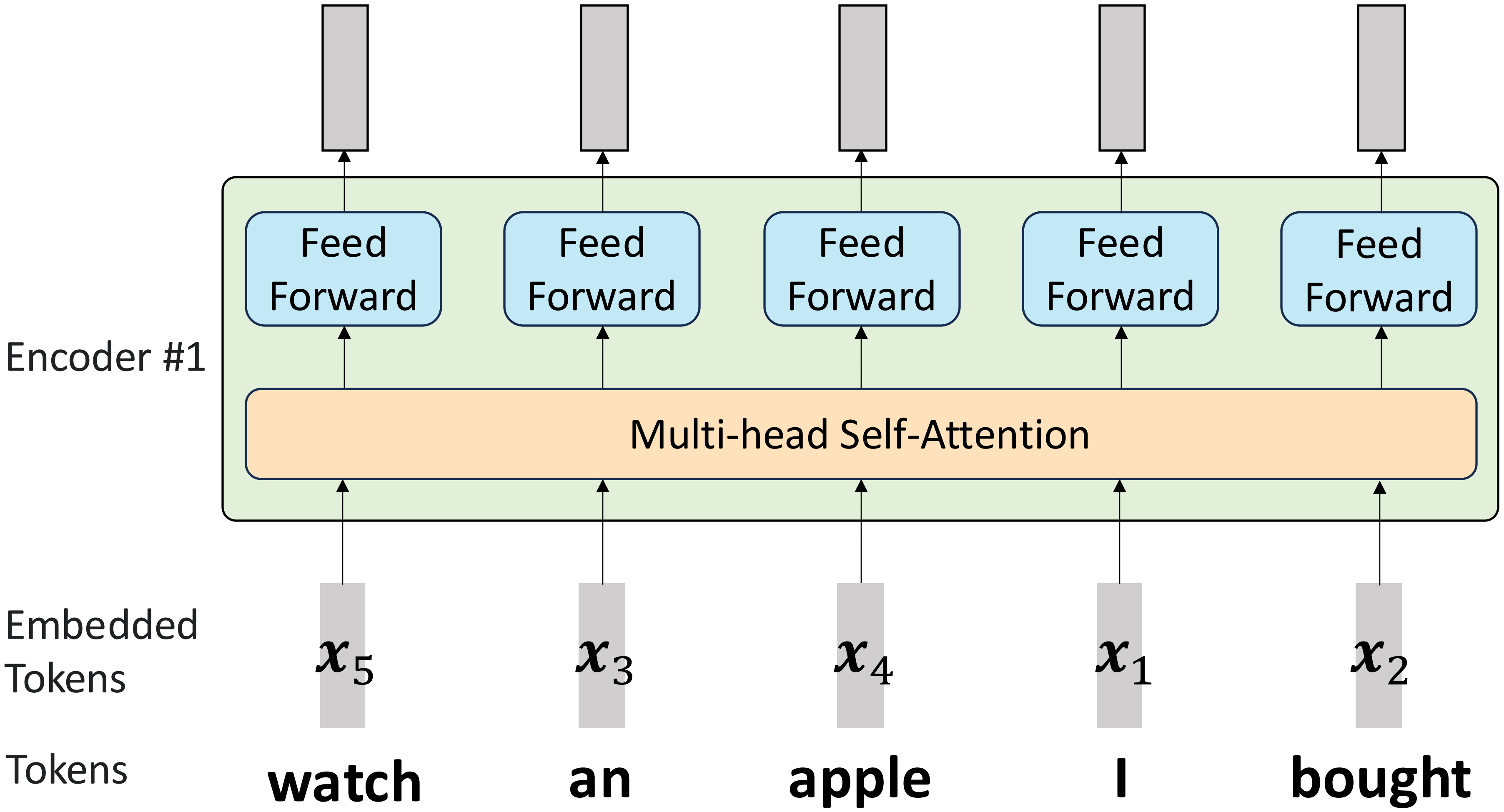
Multi-head Self-Attention

Embedded Tokens

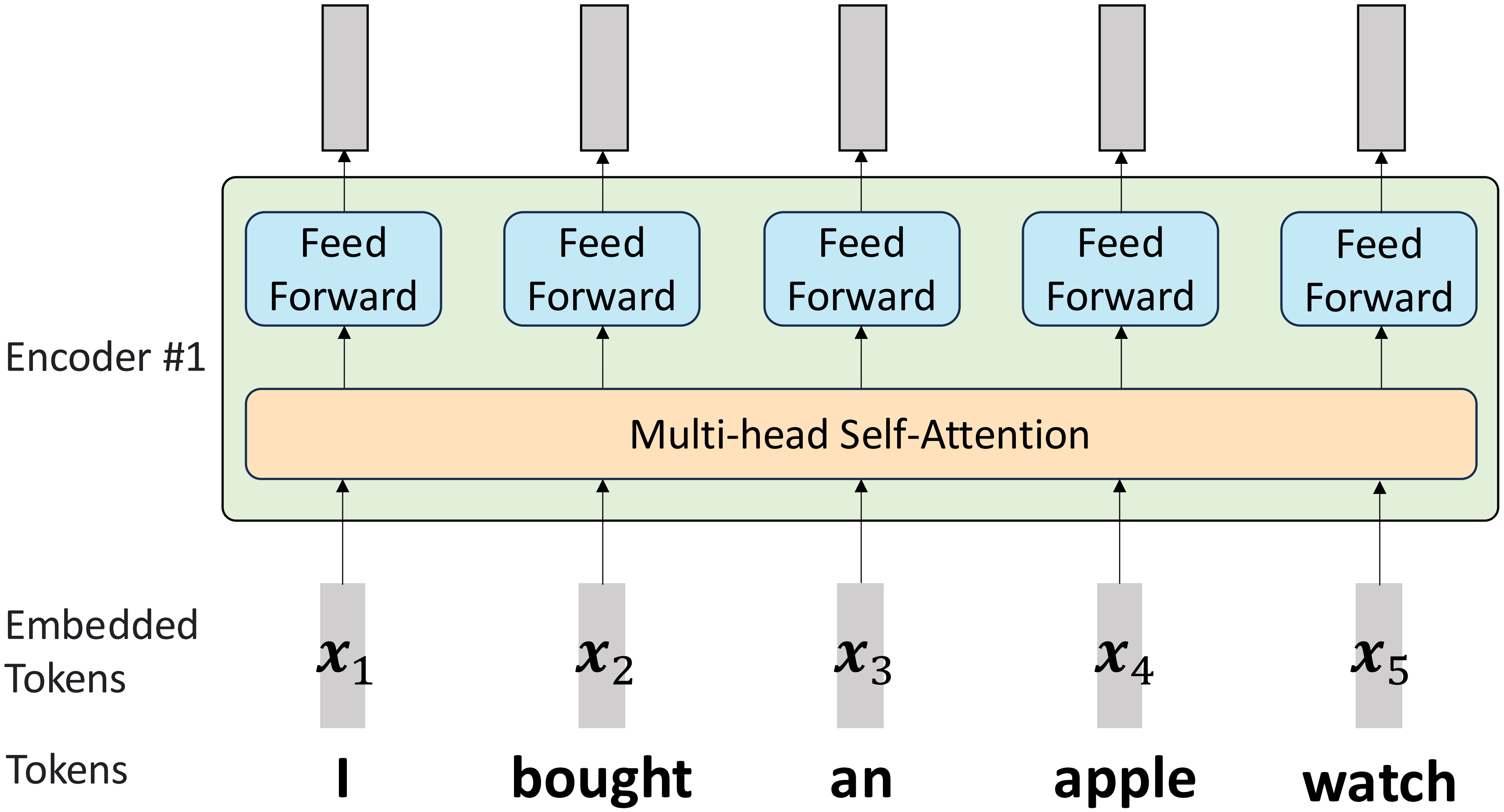
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens

I **bought** **an** **apple** **watch**

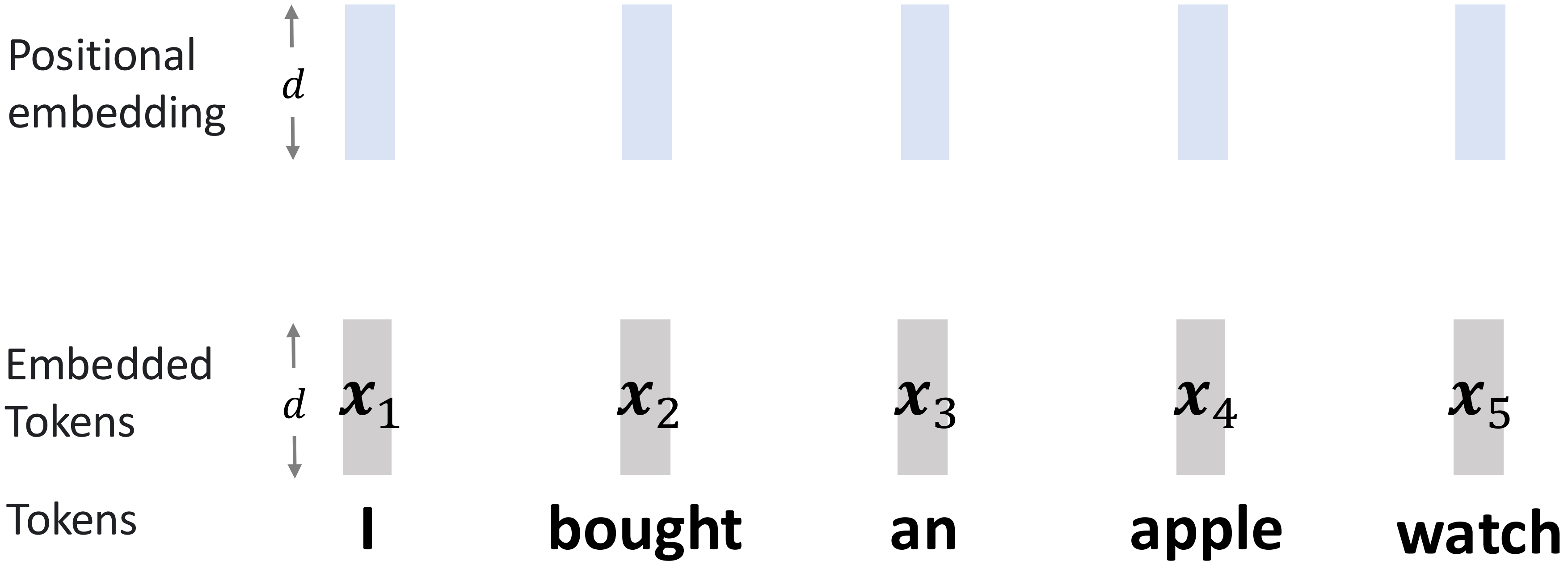

Encoder #1
Feed Forward
Feed Forward
Feed Forward
Feed Forward
Feed Forward
Multi-head Self-Attention
Embedded Tokens
x_5
x_3
x_4
x_1
x_2
Tokens
watch
an
apple
I
bought

Positional encoding



Positional encoding

	Position k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Dimension	2^3	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	← Slow oscillating
	2^2	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	
	2^1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	
	2^0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	← Fast oscillating



Positional embedding

d

Embedded Tokens

d

$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens

I **bought** **an** **apple** **watch**

Positional encoding

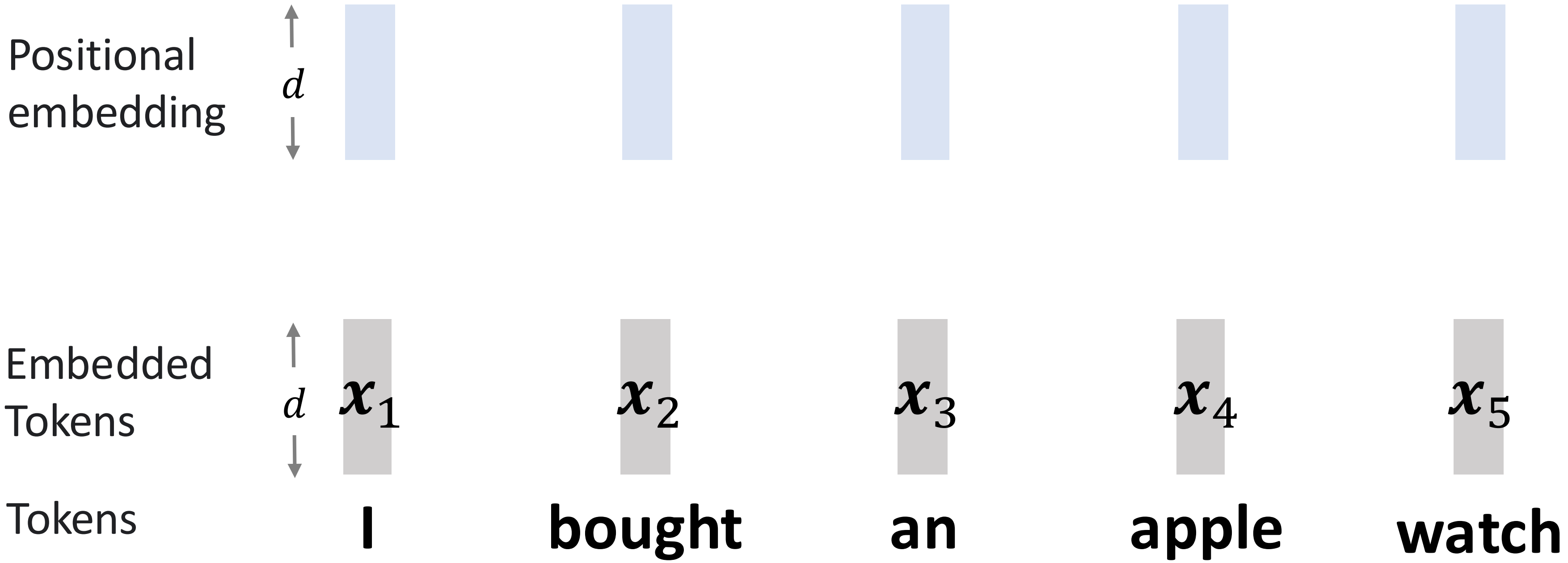
Position k

Angular frequency

$$w_i = N^{-2i/d}$$

$$N = 100{,}000$$

$$d \quad \begin{bmatrix} \sin(w_0 k) \\ \cos(w_0 k) \\ \sin(w_1 k) \\ \cos(w_1 k) \\ \vdots \\ \vdots \\ \sin\left(w_{\frac{d}{2}-1} k\right) \\ \cos\left(w_{\frac{d}{2}-1} k\right) \end{bmatrix} \begin{matrix} \longleftarrow \text{Fast oscillating} \\ \\ \\ \\ \\ \\ \longleftarrow \text{Slow oscillating} \\ \\ \end{matrix}$$



Positional encoding

- Normalized Range
- Unique identifier, unlimited length
- Relative positions as linear transform

Position k

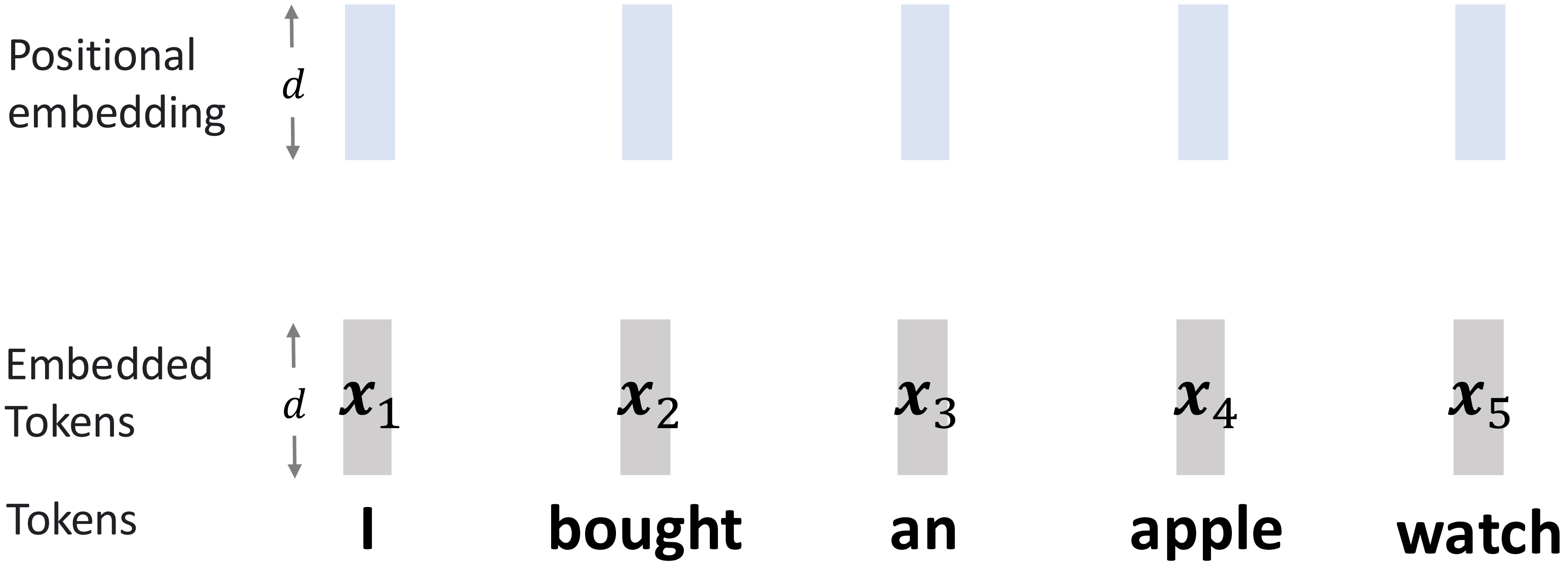
Angular frequency

$$w_i = N^{-2i/d}$$

$$N = 100{,}000$$

$$d \updownarrow \begin{bmatrix} \sin(w_0 k) \\ \cos(w_0 k) \\ \sin(w_1 k) \\ \cos(w_1 k) \\ \vdots \\ \vdots \\ \sin\left(w_{\frac{d}{2}-1} k\right) \\ \cos\left(w_{\frac{d}{2}-1} k\right) \end{bmatrix}$$

$$\begin{bmatrix} \sin(w_i(k + \Delta k)) \\ \cos(w_i(k + \Delta k)) \end{bmatrix} = \begin{bmatrix} \sin(w_i k)\cos(w_i \Delta k) + \cos(w_i k)\sin(w_i \Delta k) \\ \cos(w_i k)\cos(w_i \Delta k) - \sin(w_i k)\sin(w_i \Delta k) \end{bmatrix}$$



Positional embedding

d

Embedded Tokens

d

$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens

I **bought** **an** **apple** **watch**

Positional encoding

- Normalized Range
- Unique identifier, unlimited length
- Relative positions as linear transform

Position k

Angular frequency

$$w_i = N^{-2i/d}$$

$$N = 100{,}000$$

$$d \updownarrow \begin{bmatrix} \sin(w_0 k) \\ \cos(w_0 k) \\ \sin(w_1 k) \\ \cos(w_1 k) \\ \vdots \\ \vdots \\ \sin\left(w_{\frac{d}{2}-1} k\right) \\ \cos\left(w_{\frac{d}{2}-1} k\right) \end{bmatrix}$$

$$\begin{bmatrix} \sin(w_i(k+\Delta k)) \\ \cos(w_i(k+\Delta k)) \end{bmatrix} = \begin{bmatrix} \sin(w_i k)\cos(w_i \Delta k) + \cos(w_i k)\sin(w_i \Delta k) \\ \cos(w_i k)\cos(w_i \Delta k) - \sin(w_i k)\sin(w_i \Delta k) \end{bmatrix}$$

$$= \begin{bmatrix} \cos(w_i \Delta k) & \sin(w_i \Delta k) \\ -\sin(w_i \Delta k) & \cos(w_i \Delta k) \end{bmatrix} \begin{bmatrix} \sin(w_i k) \\ \cos(w_i k) \end{bmatrix}$$

$$P_{k+\Delta k} = M P_k$$

Positional embedding	d	$\boldsymbol{P}_1$	$\boldsymbol{P}_2$	$\boldsymbol{P}_3$	$\boldsymbol{P}_4$	$\boldsymbol{P}_5$
Embedded Tokens	d	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$
Tokens		**I**	**bought**	**an**	**apple**	**watch**

Positional encoding

Position k

Angular frequency

$w_i = N^{-2i/d}$

$N = 100{,}000$

$$d \begin{bmatrix} \sin(w_0 k) \\ \cos(w_0 k) \\ \sin(w_1 k) \\ \cos(w_1 k) \\ \vdots \\ \vdots \\ \sin\left(w_{\frac{d}{2}-1} k\right) \\ \cos\left(w_{\frac{d}{2}-1} k\right) \end{bmatrix}$$

Sinusoidal positional encoding

Relative positional encoding

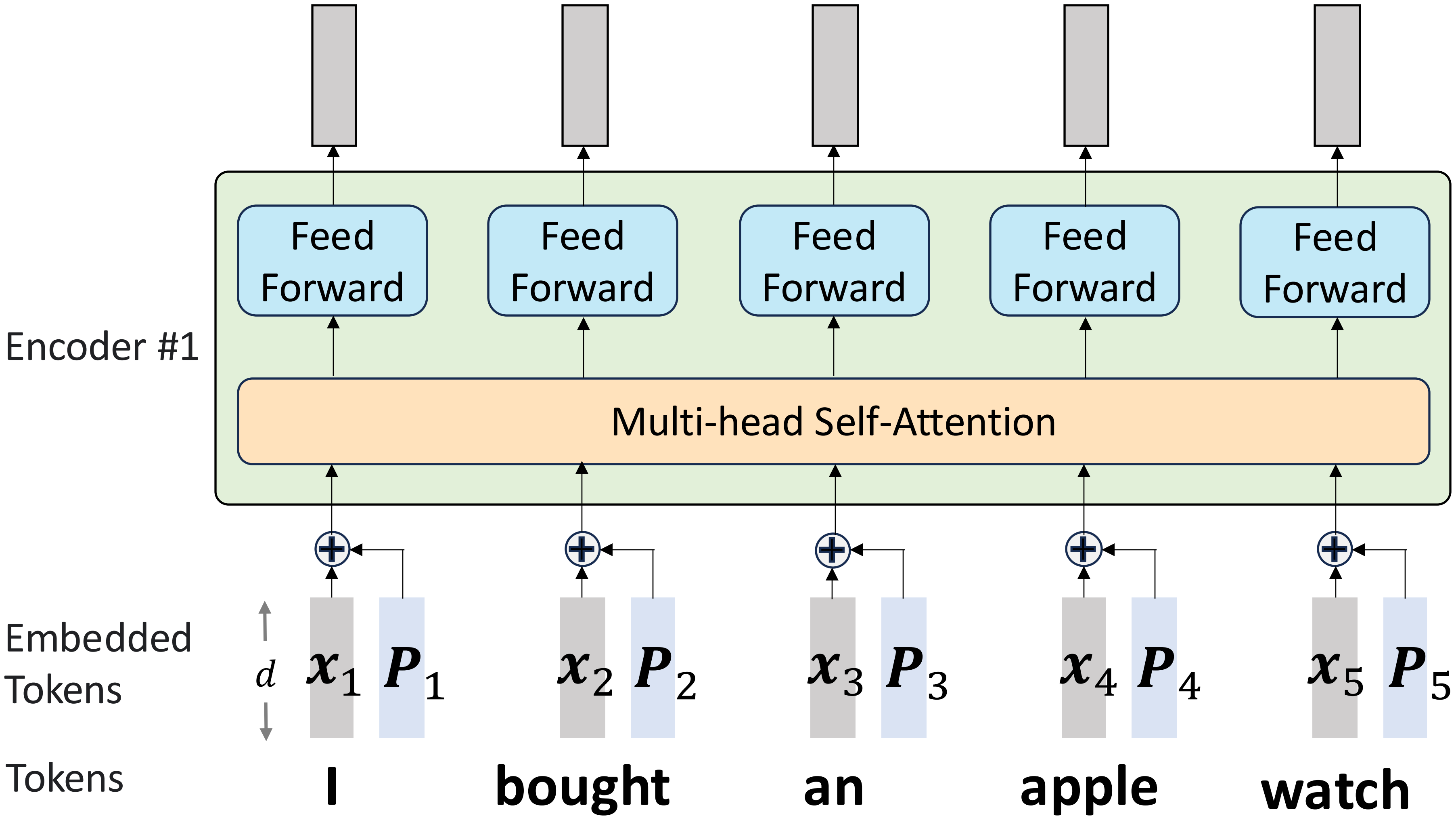
KERPLE RoPE CoPE

NoPE YaRN FIRE

Positional embedding d: $\boldsymbol{P}_1$ $\boldsymbol{P}_2$ $\boldsymbol{P}_3$ $\boldsymbol{P}_4$ $\boldsymbol{P}_5$

Embedded Tokens d: $\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$

Tokens: **I** **bought** **an** **apple** **watch**



Encoder #1

Feed Forward

Feed Forward

Feed Forward

Feed Forward

Feed Forward

Multi-head Self-Attention

Embedded Tokens

d

$\boldsymbol{x}_1$ $\boldsymbol{P}_1$

$\boldsymbol{x}_2$ $\boldsymbol{P}_2$

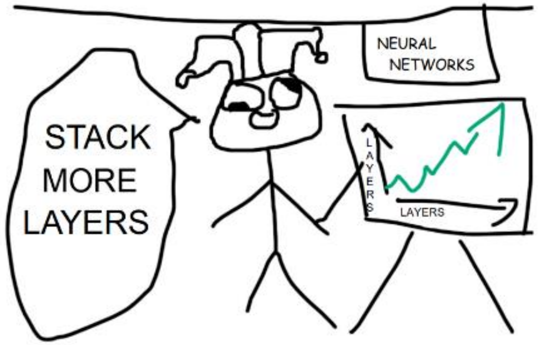
$\boldsymbol{x}_3$ $\boldsymbol{P}_3$

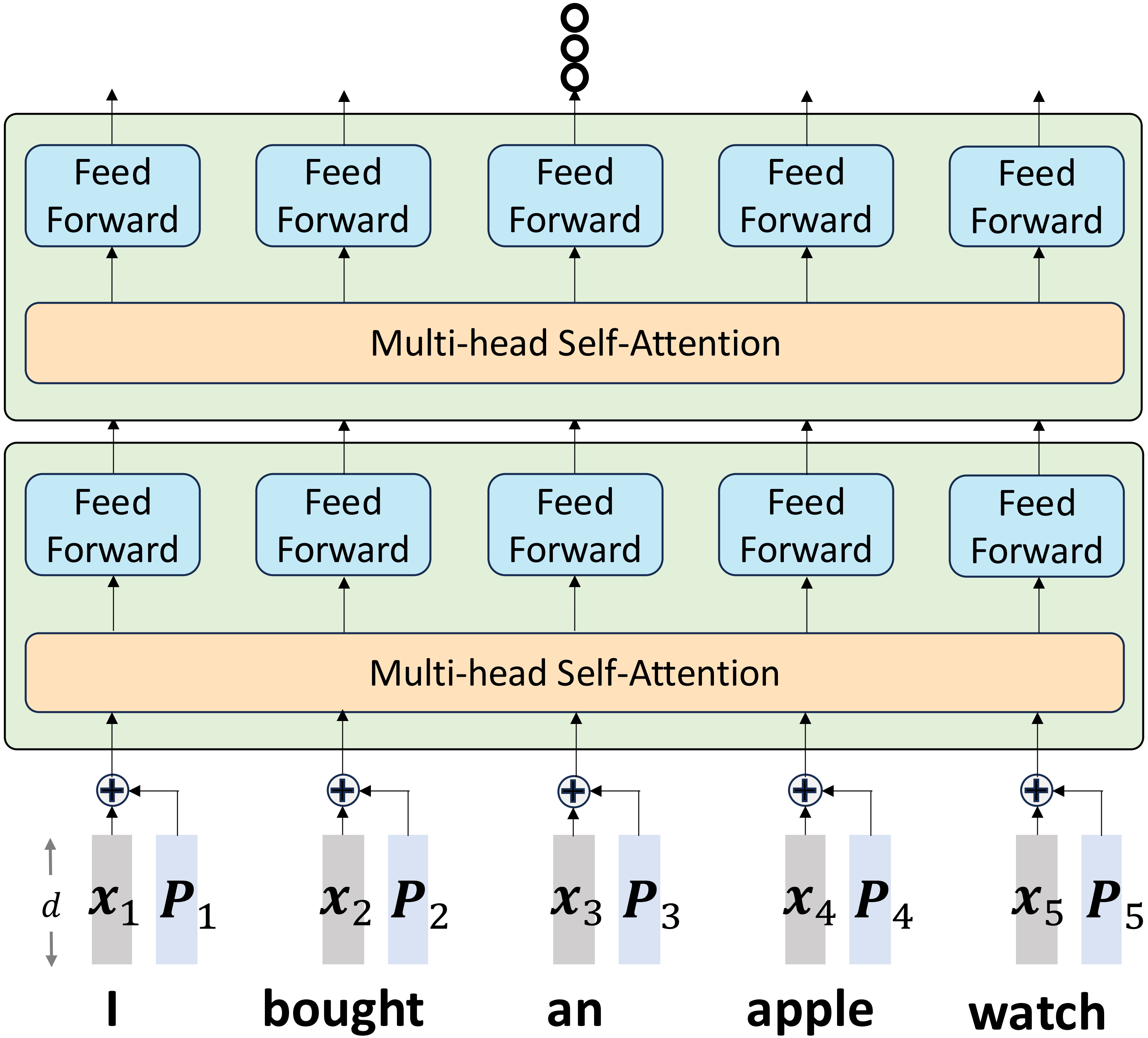
$\boldsymbol{x}_4$ $\boldsymbol{P}_4$

$\boldsymbol{x}_5$ $\boldsymbol{P}_5$

Tokens

I **bought** **an** **apple** **watch**





Encoder #2

Feed Forward | Feed Forward | Feed Forward | Feed Forward | Feed Forward

Multi-head Self-Attention

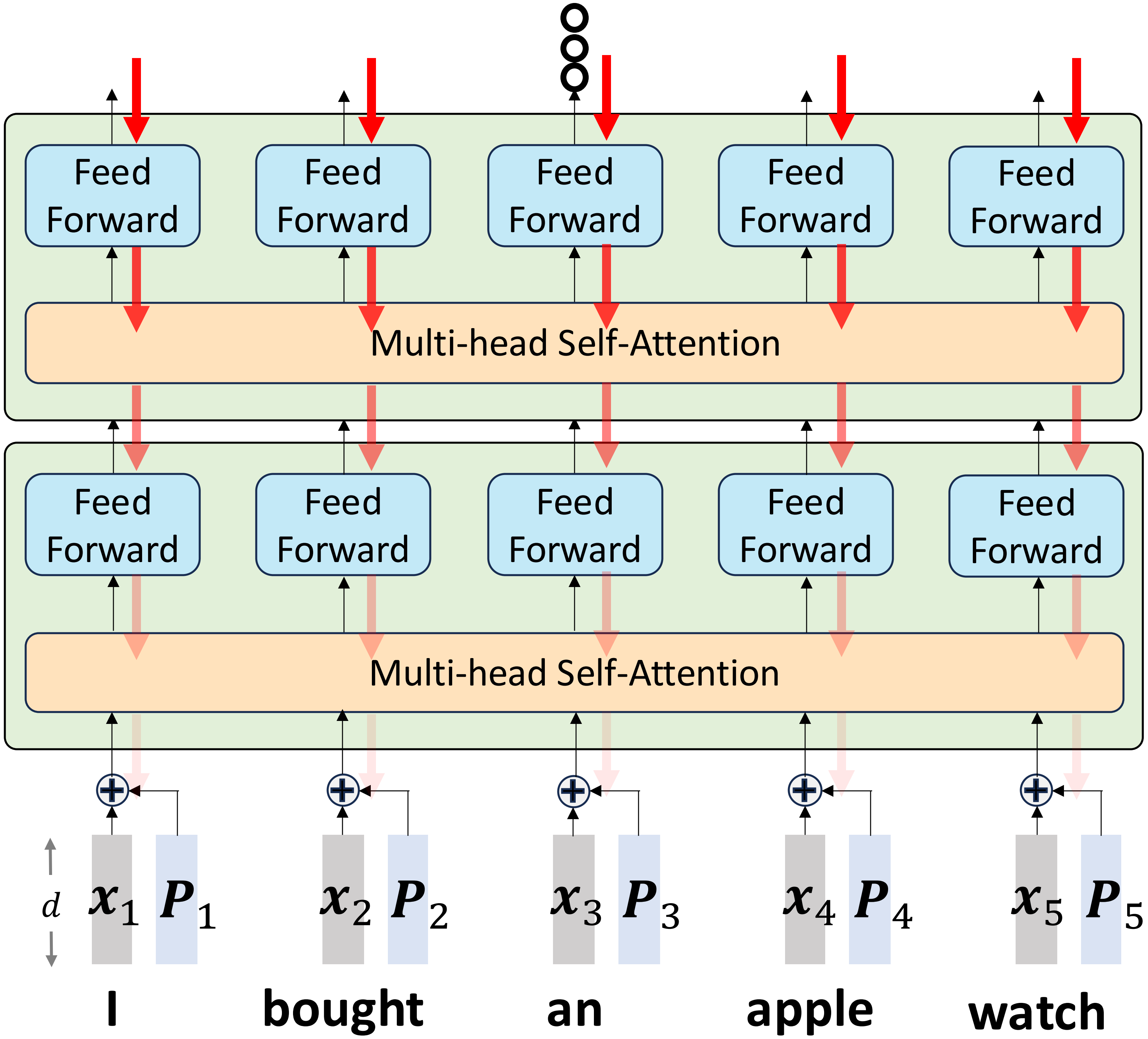
Encoder #1

Feed Forward | Feed Forward | Feed Forward | Feed Forward | Feed Forward

Multi-head Self-Attention

Embedded Tokens d $\boldsymbol{x}_1$ $\boldsymbol{P}_1$ $\boldsymbol{x}_2$ $\boldsymbol{P}_2$ $\boldsymbol{x}_3$ $\boldsymbol{P}_3$ $\boldsymbol{x}_4$ $\boldsymbol{P}_4$ $\boldsymbol{x}_5$ $\boldsymbol{P}_5$

Tokens **I** **bought** **an** **apple** **watch**



Encoder #2

Feed Forward | Feed Forward | Feed Forward | Feed Forward | Feed Forward

Multi-head Self-Attention

Encoder #1

Feed Forward | Feed Forward | Feed Forward | Feed Forward | Feed Forward

Multi-head Self-Attention

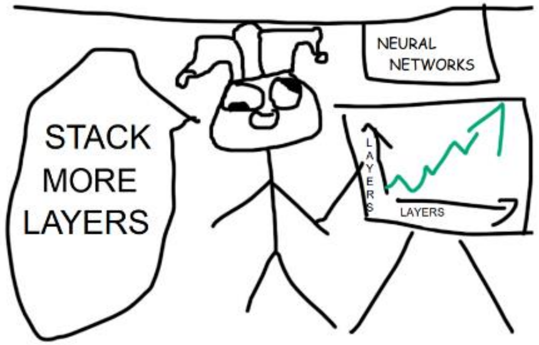
Embedded Tokens

d

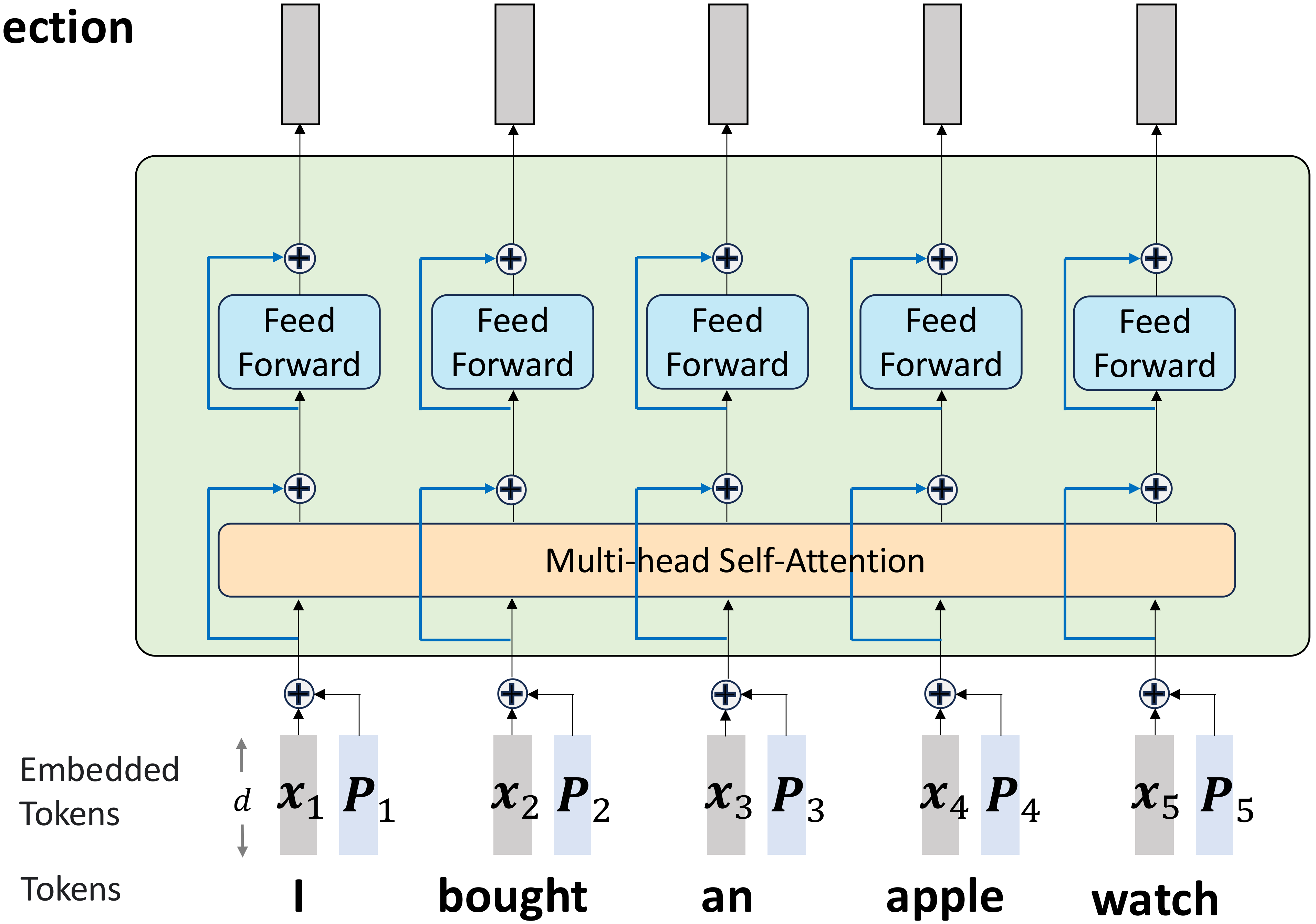
$\boldsymbol{x}_1$ $\boldsymbol{P}_1$ $\boldsymbol{x}_2$ $\boldsymbol{P}_2$ $\boldsymbol{x}_3$ $\boldsymbol{P}_3$ $\boldsymbol{x}_4$ $\boldsymbol{P}_4$ $\boldsymbol{x}_5$ $\boldsymbol{P}_5$

Tokens

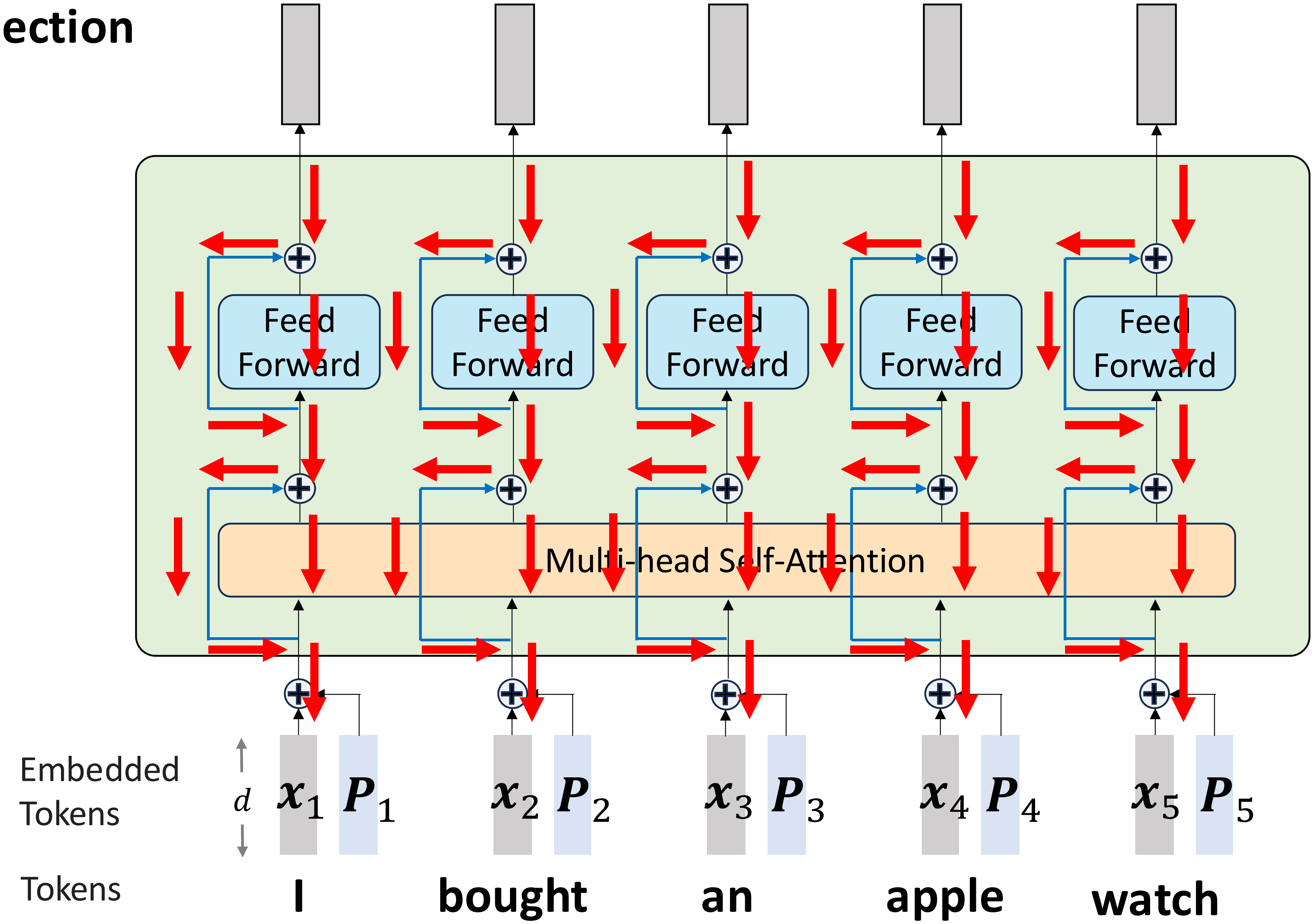
I **bought** **an** **apple** **watch**



Residual connection

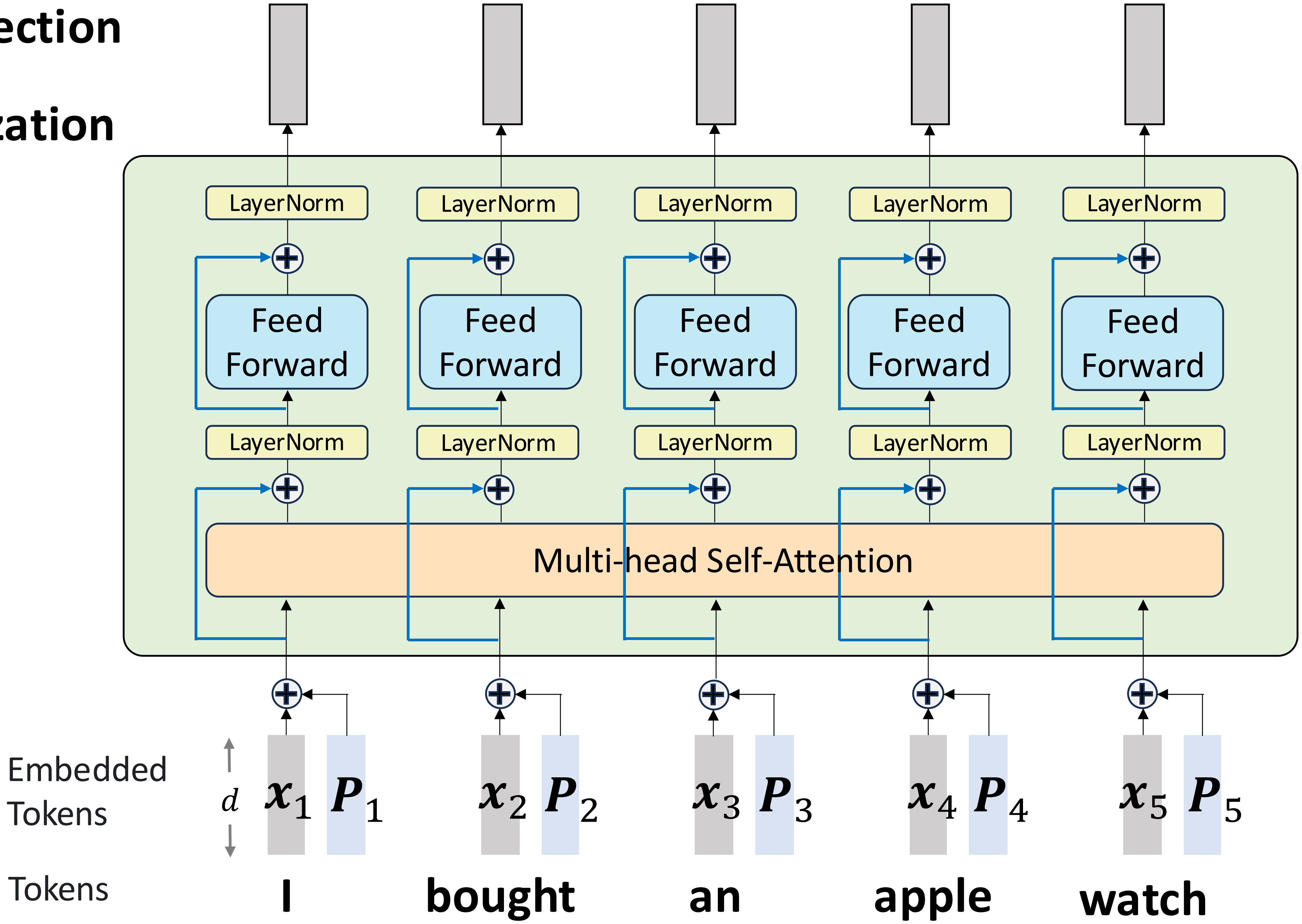


Residual connection



Residual connection

Layer normalization



Residual connection

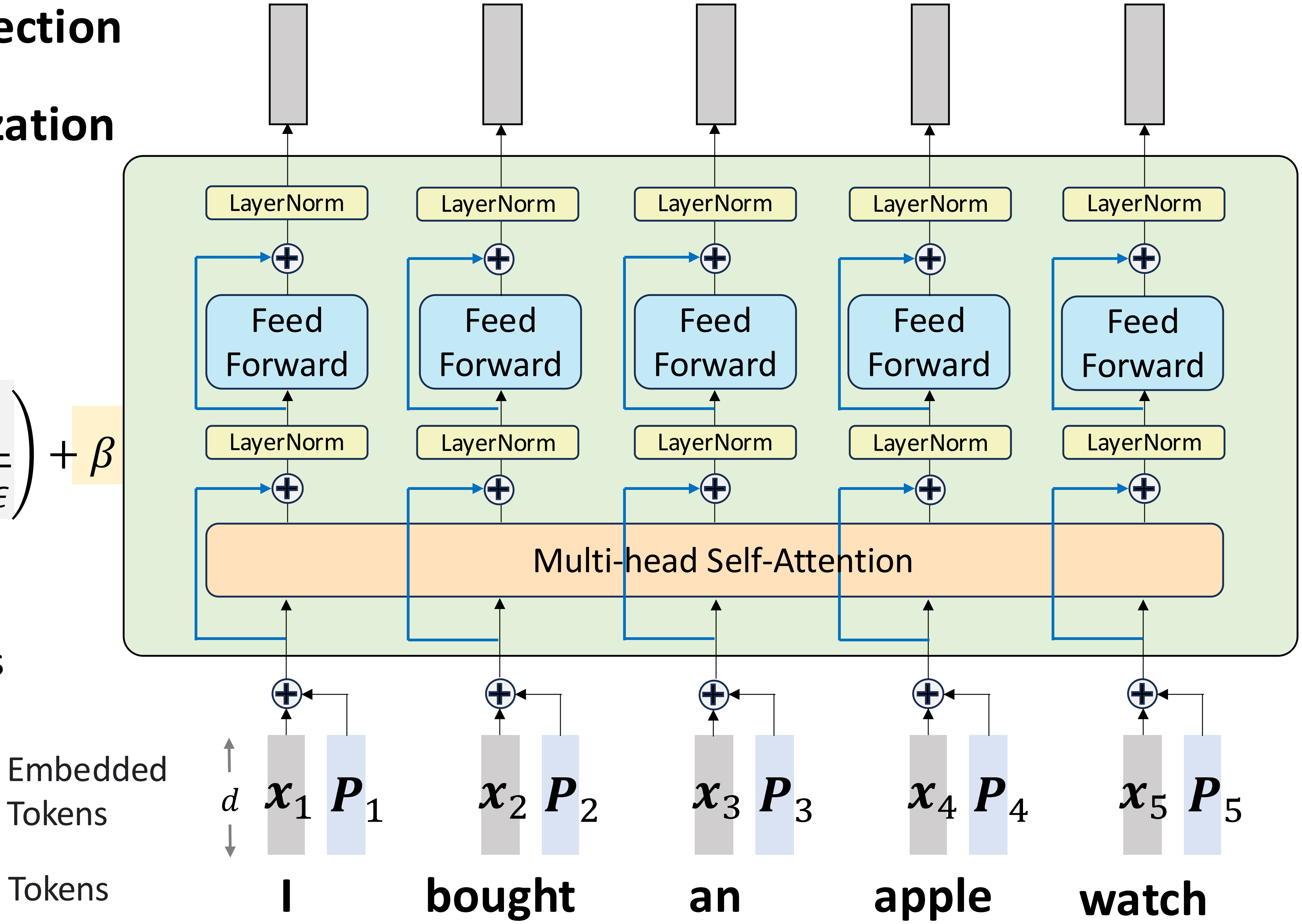
Layer normalization

$$\mathrm{LayerNorm}(\boldsymbol{x}) =$$

$$\gamma\left(\frac{\boldsymbol{x} - \mathrm{mean}(\boldsymbol{x})}{\sqrt{\mathrm{Variance}(\mathbf{x}) + \epsilon}}\right) + \beta$$

$\gamma, \beta \in R$

Learnable parameters



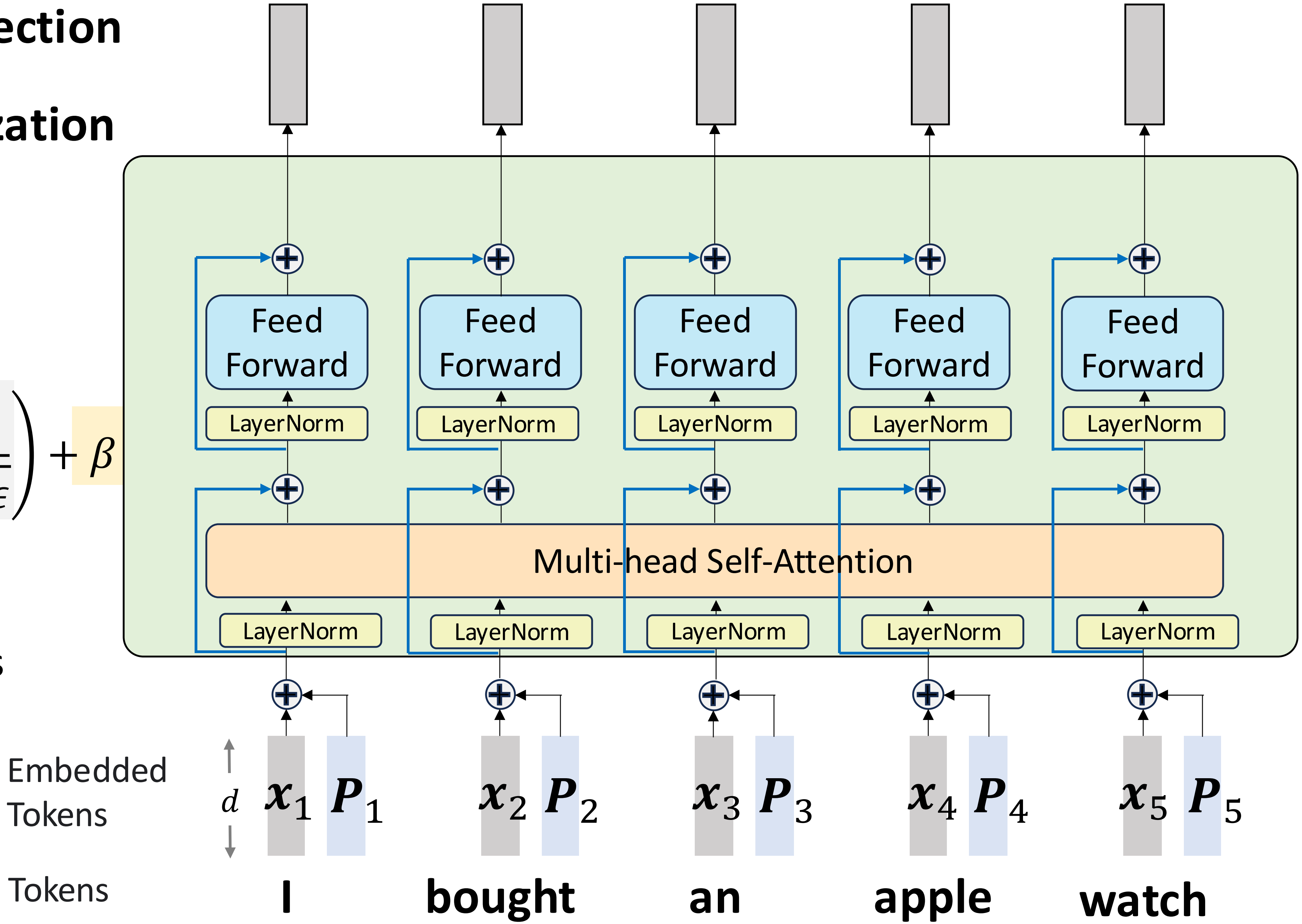
Residual connection

Layer normalization

$$\mathrm{LayerNorm}(\boldsymbol{x}) = \gamma\left(\frac{\boldsymbol{x} - \mathrm{mean}(\boldsymbol{x})}{\sqrt{\mathrm{Variance}(\mathbf{x}) + \epsilon}}\right) + \beta$$

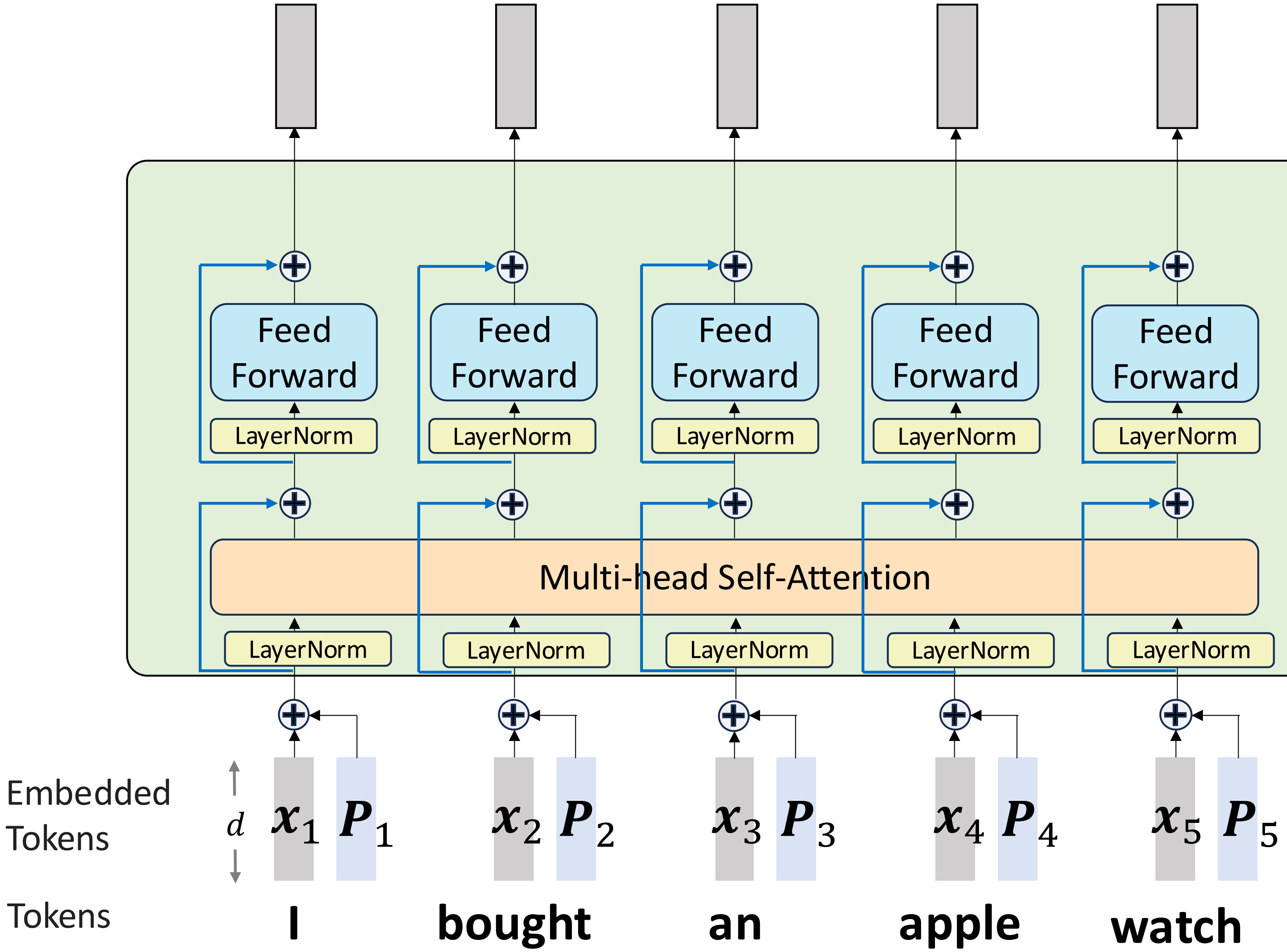
$\gamma, \beta \in R$

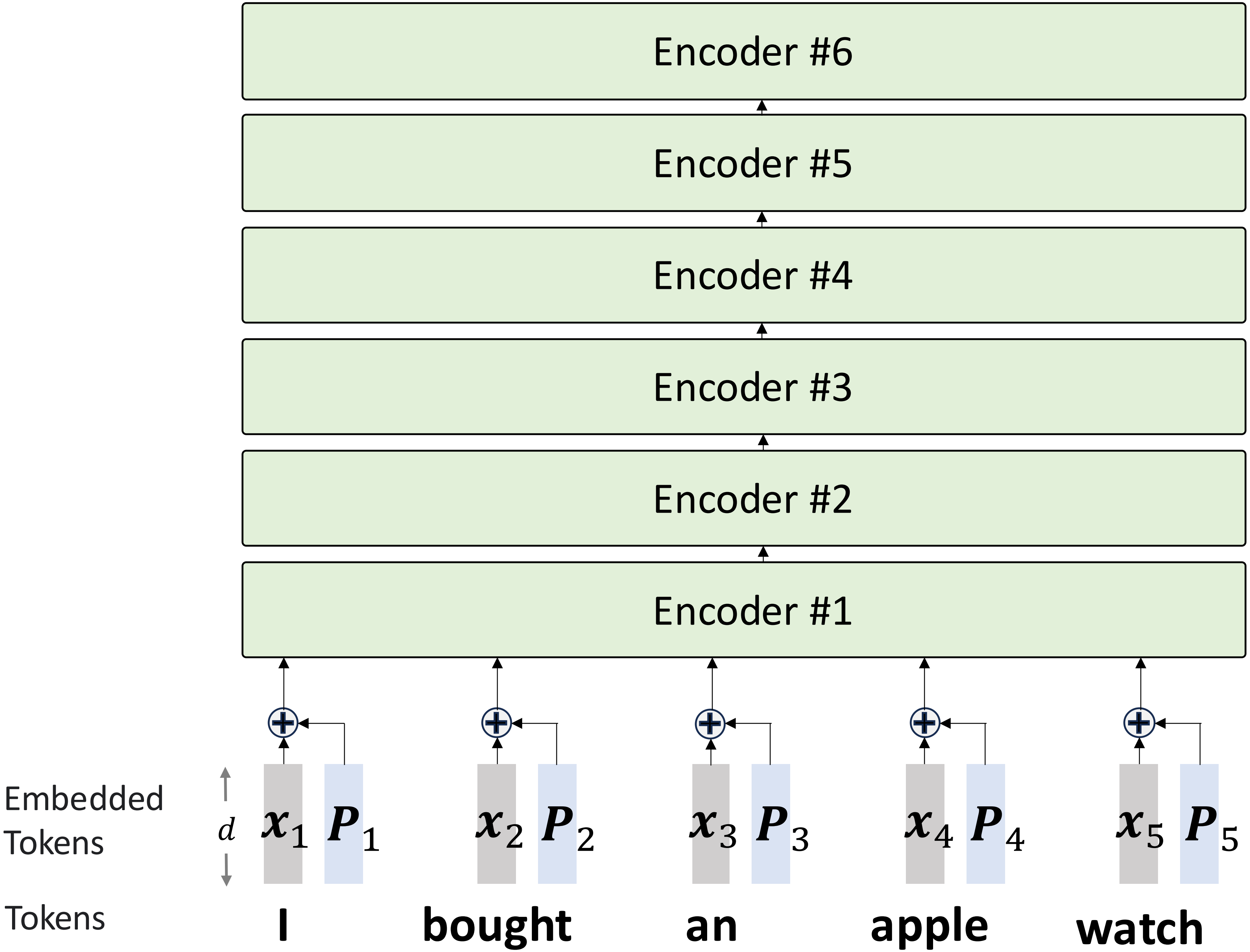
Learnable parameters


Feed Forward
LayerNorm
Multi-head Self-Attention
LayerNorm
Embedded Tokens
Tokens
d
$\boldsymbol{x}_1$ $\boldsymbol{P}_1$
$\boldsymbol{x}_2$ $\boldsymbol{P}_2$
$\boldsymbol{x}_3$ $\boldsymbol{P}_3$
$\boldsymbol{x}_4$ $\boldsymbol{P}_4$
$\boldsymbol{x}_5$ $\boldsymbol{P}_5$
I
bought
an
apple
watch

[Xiong et al. 2020]

On Layer Normalization in the Transformer Architecture


Feed Forward
LayerNorm
Multi-head Self-Attention
LayerNorm
Embedded Tokens
d
x_1 P_1
x_2 P_2
x_3 P_3
x_4 P_4
x_5 P_5
Tokens
I
bought
an
apple
watch



Encoder #6

Encoder #5

Encoder #4

Encoder #3

Encoder #2

Encoder #1

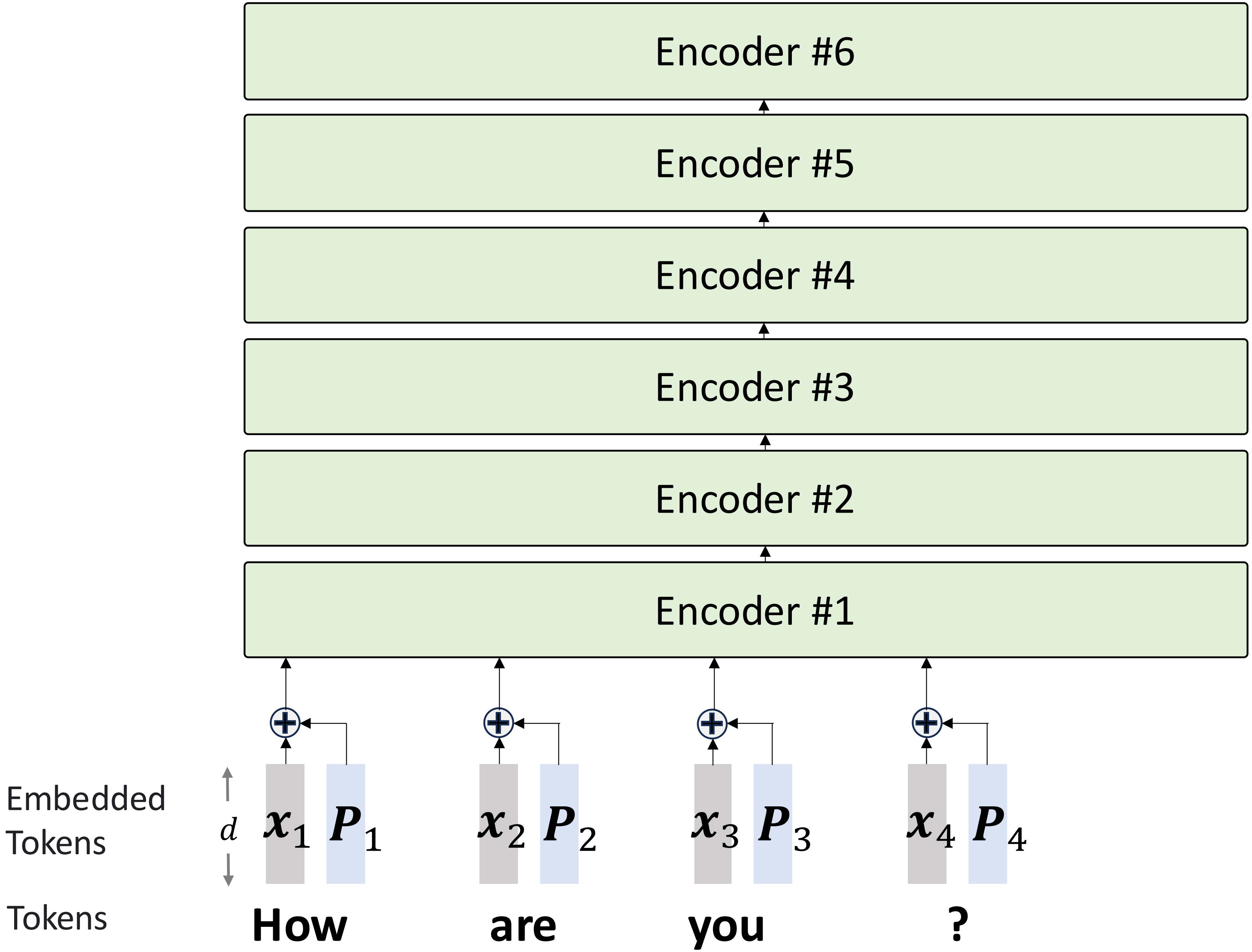
Embedded Tokens

d

$\boldsymbol{x}_1$ $\boldsymbol{P}_1$ $\boldsymbol{x}_2$ $\boldsymbol{P}_2$ $\boldsymbol{x}_3$ $\boldsymbol{P}_3$ $\boldsymbol{x}_4$ $\boldsymbol{P}_4$ $\boldsymbol{x}_5$ $\boldsymbol{P}_5$

Tokens

I **bought** **an** **apple** **watch**



Encoder #6

Encoder #5

Encoder #4

Encoder #3

Encoder #2

Encoder #1

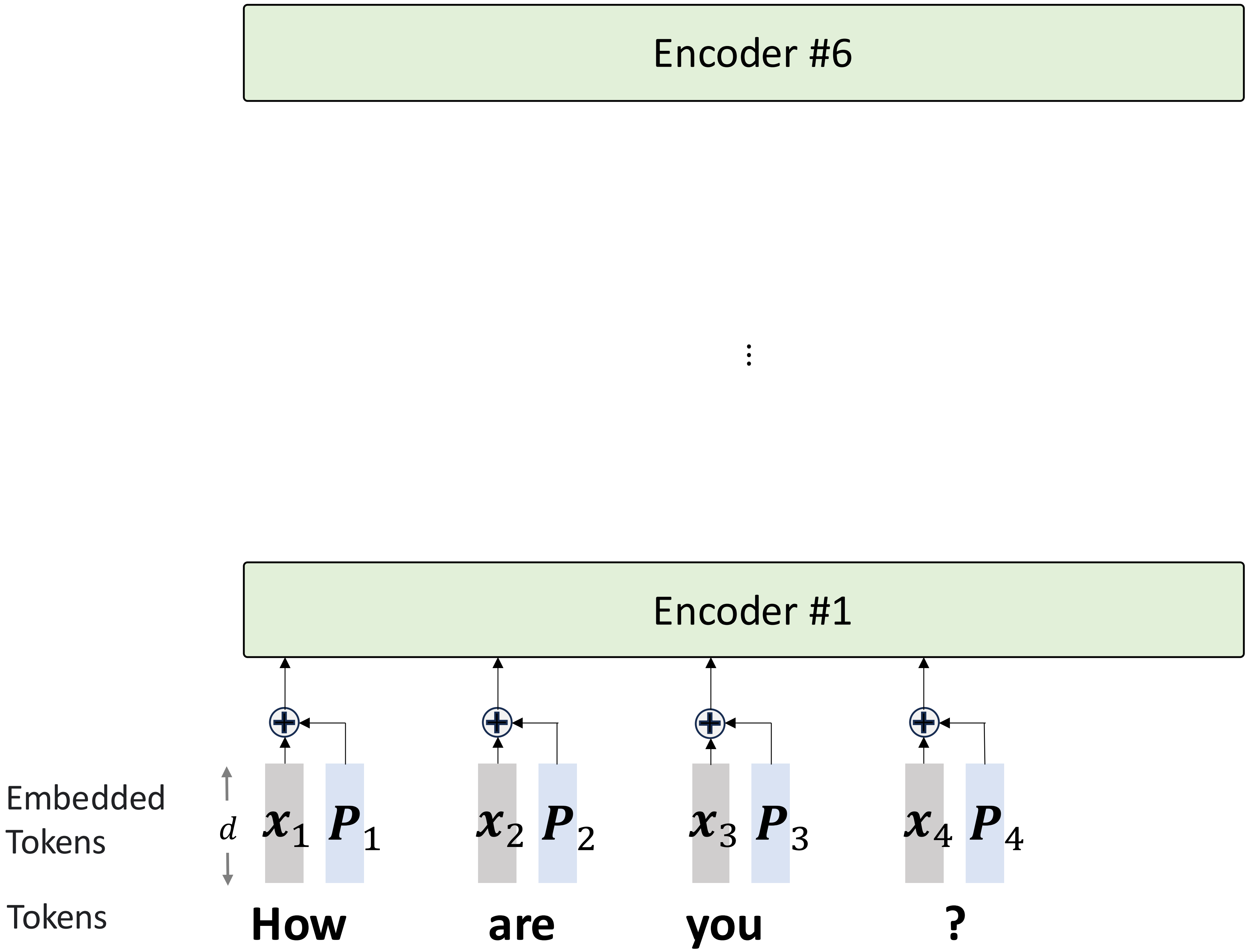
Embedded Tokens

d

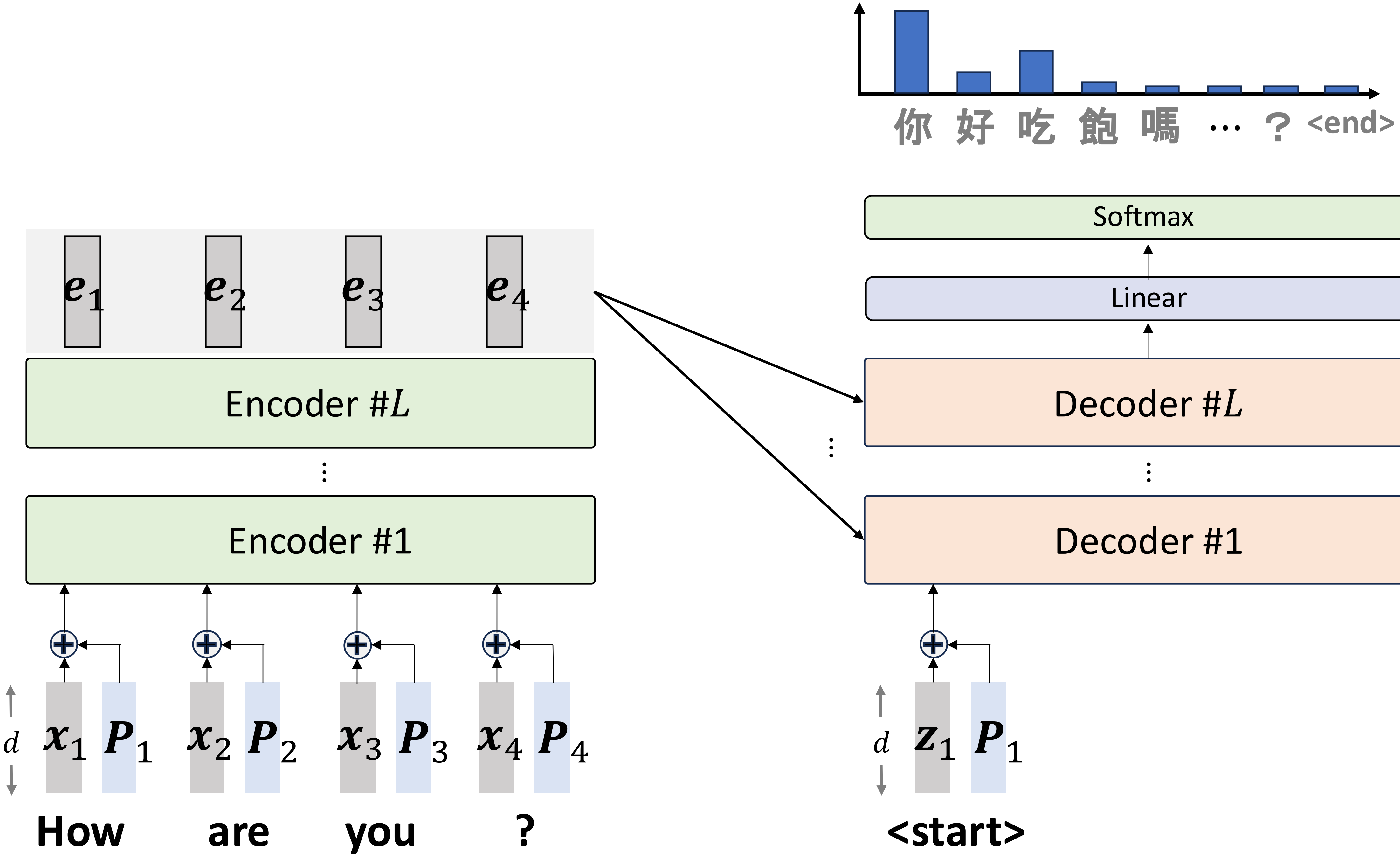
$\boldsymbol{x}_1$ $\boldsymbol{P}_1$ $\boldsymbol{x}_2$ $\boldsymbol{P}_2$ $\boldsymbol{x}_3$ $\boldsymbol{P}_3$ $\boldsymbol{x}_4$ $\boldsymbol{P}_4$

Tokens

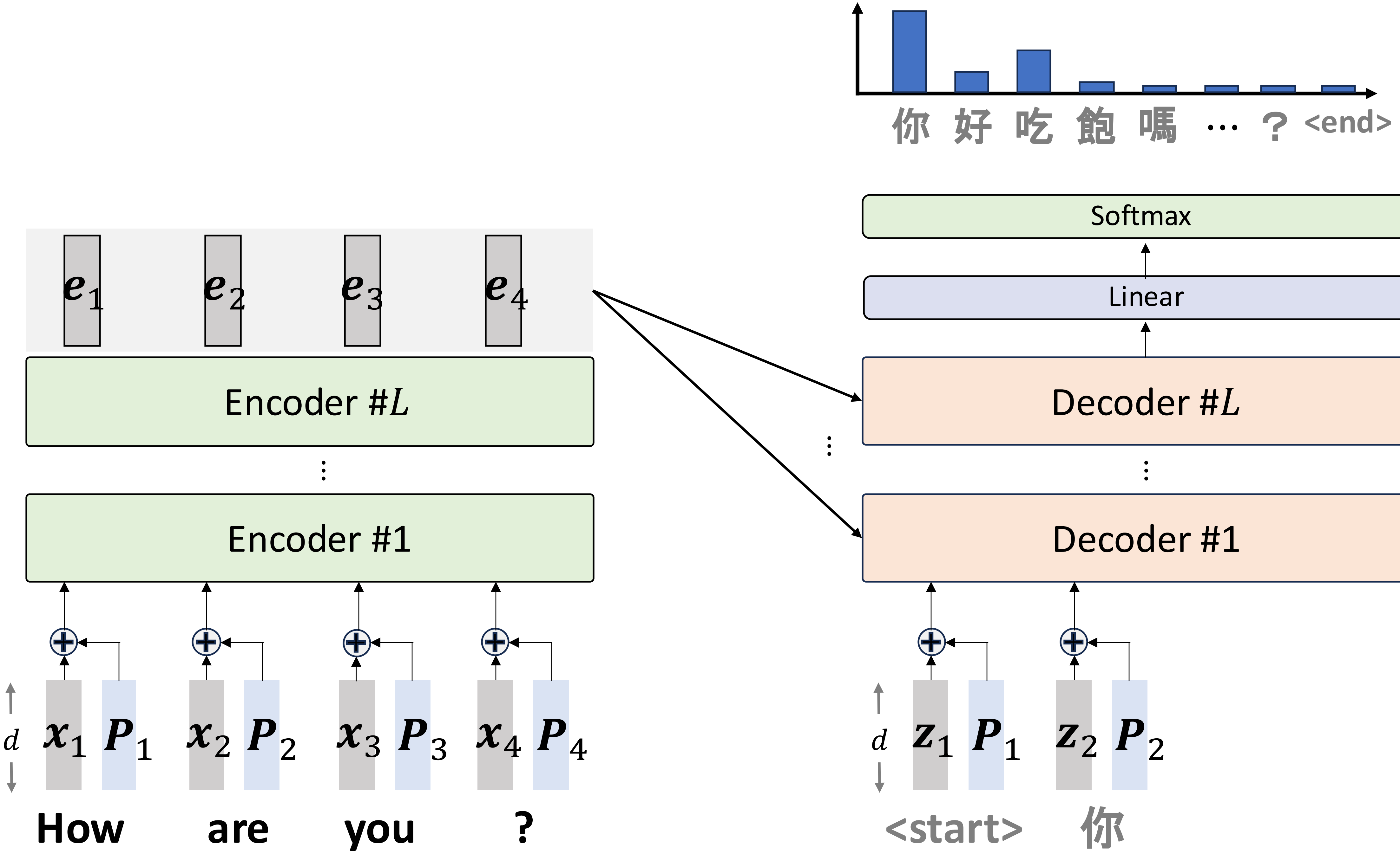
How **are** **you** **?**



Encoder #6
⋮
Encoder #1
Embedded Tokens
d
$\boldsymbol{x}_1$ $\boldsymbol{P}_1$
$\boldsymbol{x}_2$ $\boldsymbol{P}_2$
$\boldsymbol{x}_3$ $\boldsymbol{P}_3$
$\boldsymbol{x}_4$ $\boldsymbol{P}_4$
Tokens
How
are
you
?



你
好
吃
飽
嗎
…
？
<end>
Softmax
Linear
e_1
e_2
e_3
e_4
Encoder #L
Decoder #L
⋮
Encoder #1
Decoder #1
d
x_1 P_1
x_2 P_2
x_3 P_3
x_4 P_4
How
are
you
?
d
z_1 P_1
<start>



你 好 吃 飽 嗎 … ？ <end>

Softmax

Linear

e_1 e_2 e_3 e_4

Encoder #L

⋮

Encoder #1

Decoder #L

⋮

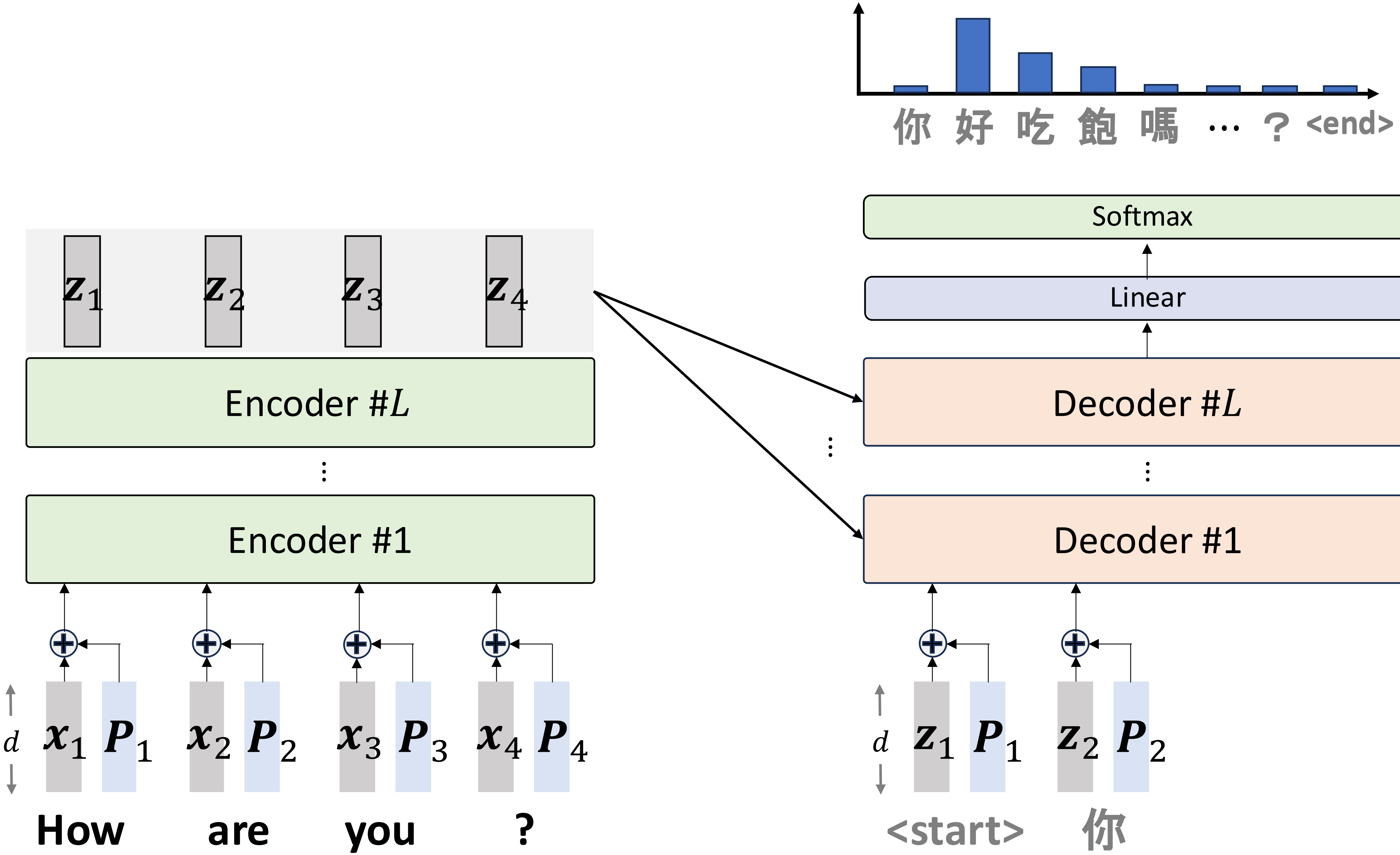
Decoder #1

d x_1 P_1 x_2 P_2 x_3 P_3 x_4 P_4

How are you ?

d z_1 P_1 z_2 P_2

<start> 你



你 好 吃 飽 嗎 … ？ <end>

Softmax

Linear

$\boldsymbol{z}_1$ $\boldsymbol{z}_2$ $\boldsymbol{z}_3$ $\boldsymbol{z}_4$

Encoder #L

Decoder #L

Encoder #1

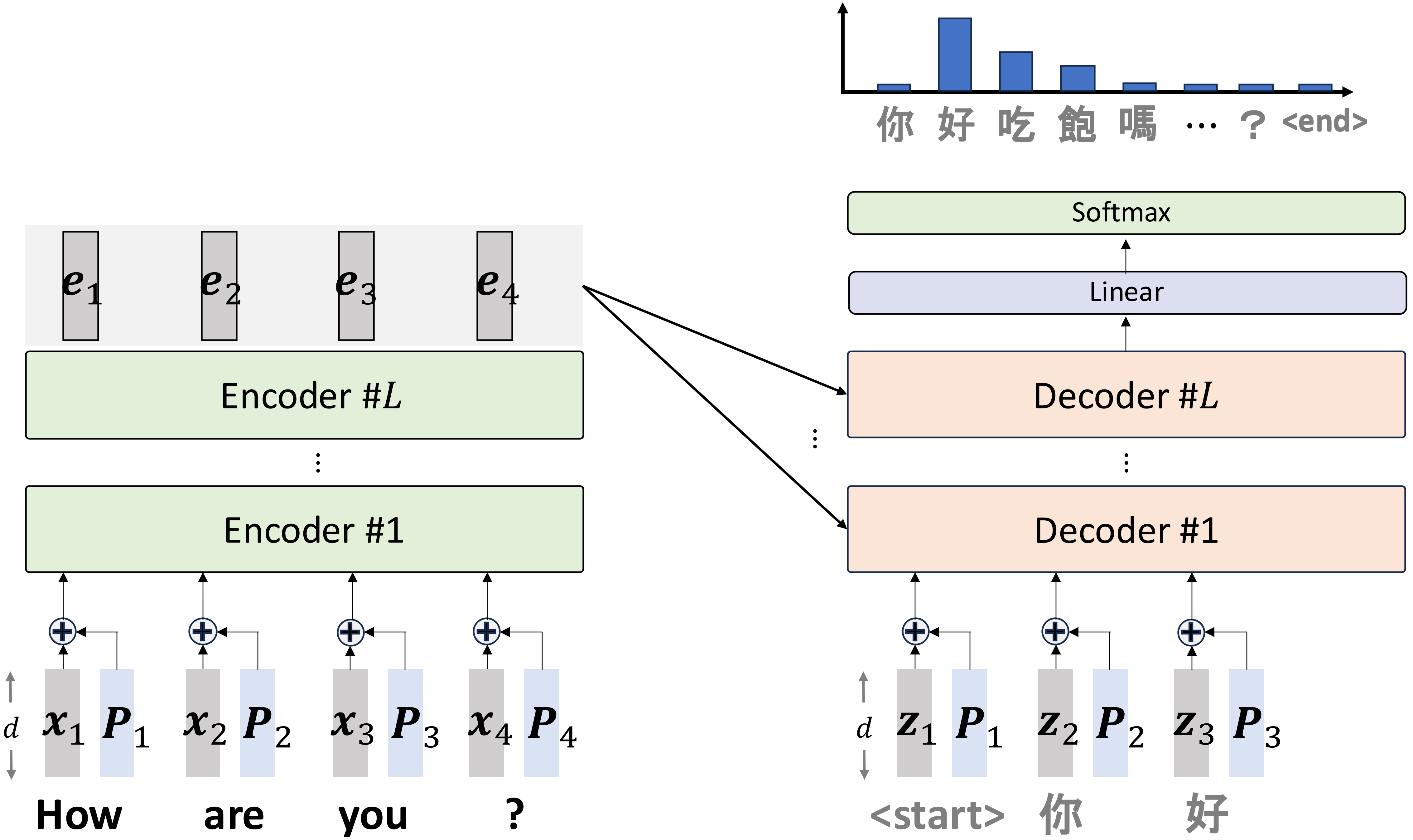
Decoder #1

d $\boldsymbol{x}_1$ $\boldsymbol{P}_1$ $\boldsymbol{x}_2$ $\boldsymbol{P}_2$ $\boldsymbol{x}_3$ $\boldsymbol{P}_3$ $\boldsymbol{x}_4$ $\boldsymbol{P}_4$

How **are** **you** **?**

d $\boldsymbol{z}_1$ $\boldsymbol{P}_1$ $\boldsymbol{z}_2$ $\boldsymbol{P}_2$

<start> **你**



你 好 吃 飽 嗎 … ？ <end>

Softmax

Linear

Decoder #L

Decoder #1

e_1 e_2 e_3 e_4

Encoder #L

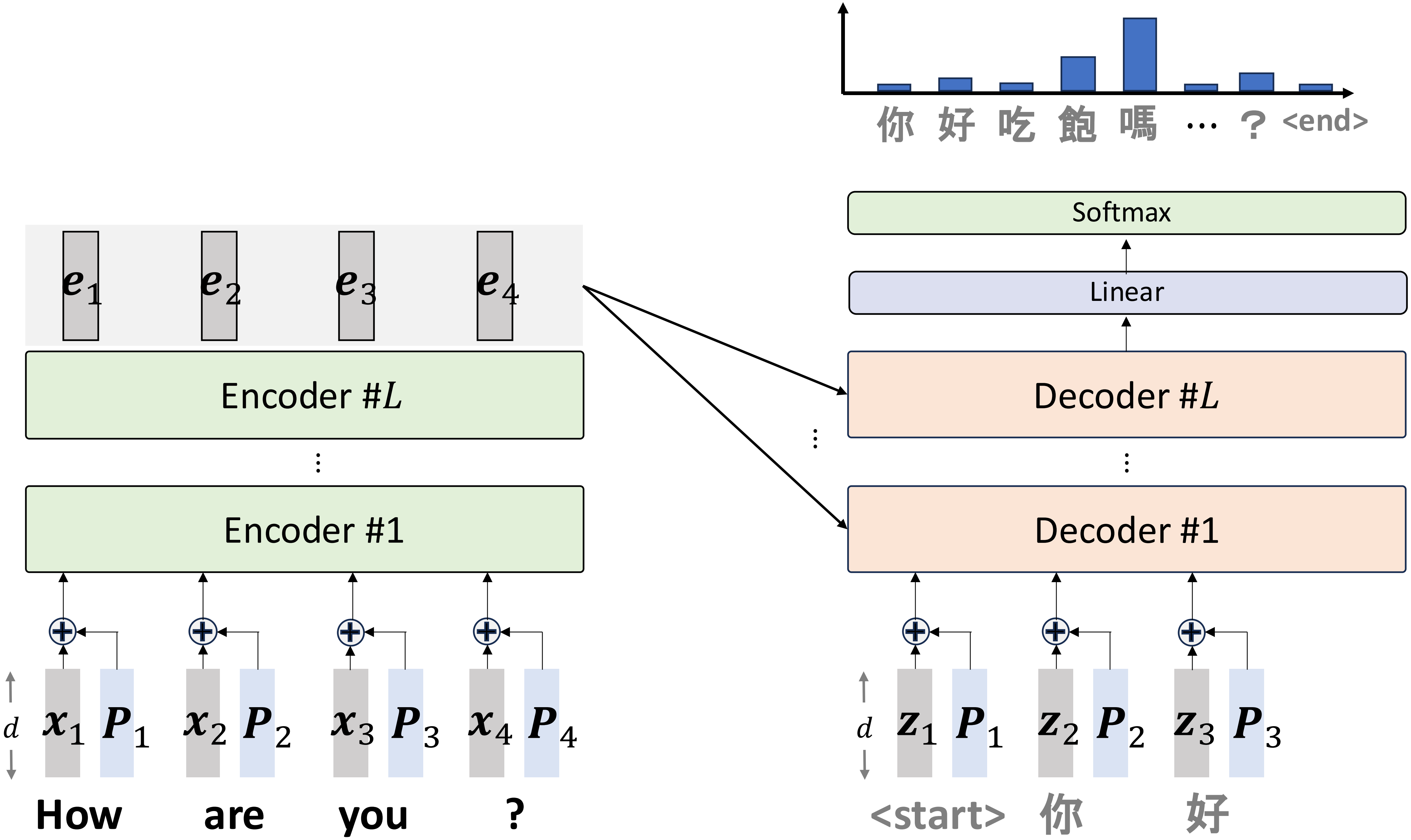
Encoder #1

d x_1 P_1 x_2 P_2 x_3 P_3 x_4 P_4

How are you ?

d z_1 P_1 z_2 P_2 z_3 P_3

<start> 你 好



你 好 吃 飽 嗎 … ？ <end>

Softmax

Linear

Decoder #L

Decoder #1

d $\boldsymbol{z}_1$ $\boldsymbol{P}_1$ $\boldsymbol{z}_2$ $\boldsymbol{P}_2$ $\boldsymbol{z}_3$ $\boldsymbol{P}_3$

<start> 你 好

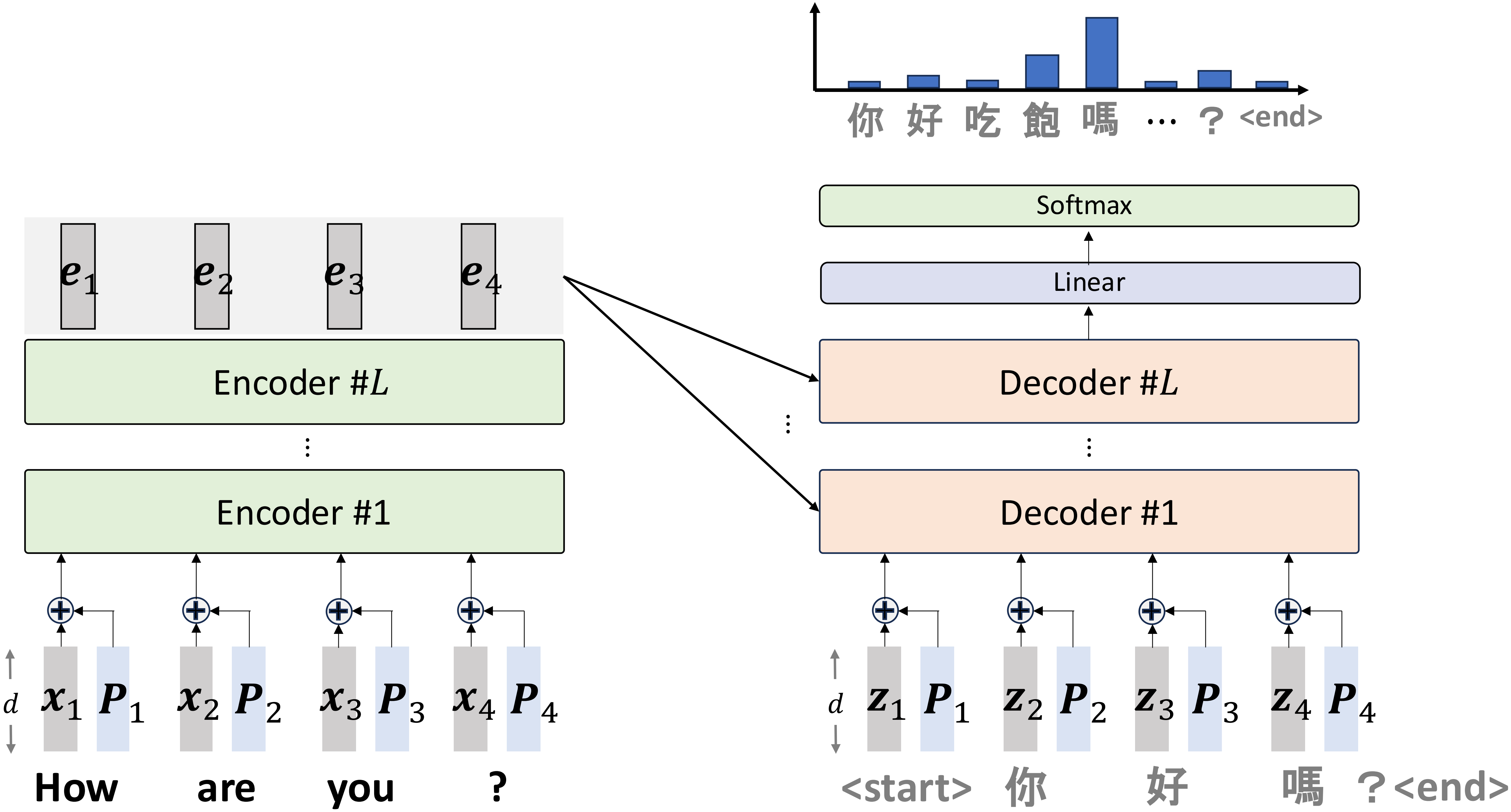
$\boldsymbol{e}_1$ $\boldsymbol{e}_2$ $\boldsymbol{e}_3$ $\boldsymbol{e}_4$

Encoder #L

Encoder #1

d $\boldsymbol{x}_1$ $\boldsymbol{P}_1$ $\boldsymbol{x}_2$ $\boldsymbol{P}_2$ $\boldsymbol{x}_3$ $\boldsymbol{P}_3$ $\boldsymbol{x}_4$ $\boldsymbol{P}_4$

How are you ?



你 好 吃 飽 嗎 … ？ <end>

Softmax

Linear

e_1 e_2 e_3 e_4

Encoder #L

⋮

Encoder #1

Decoder #L

⋮

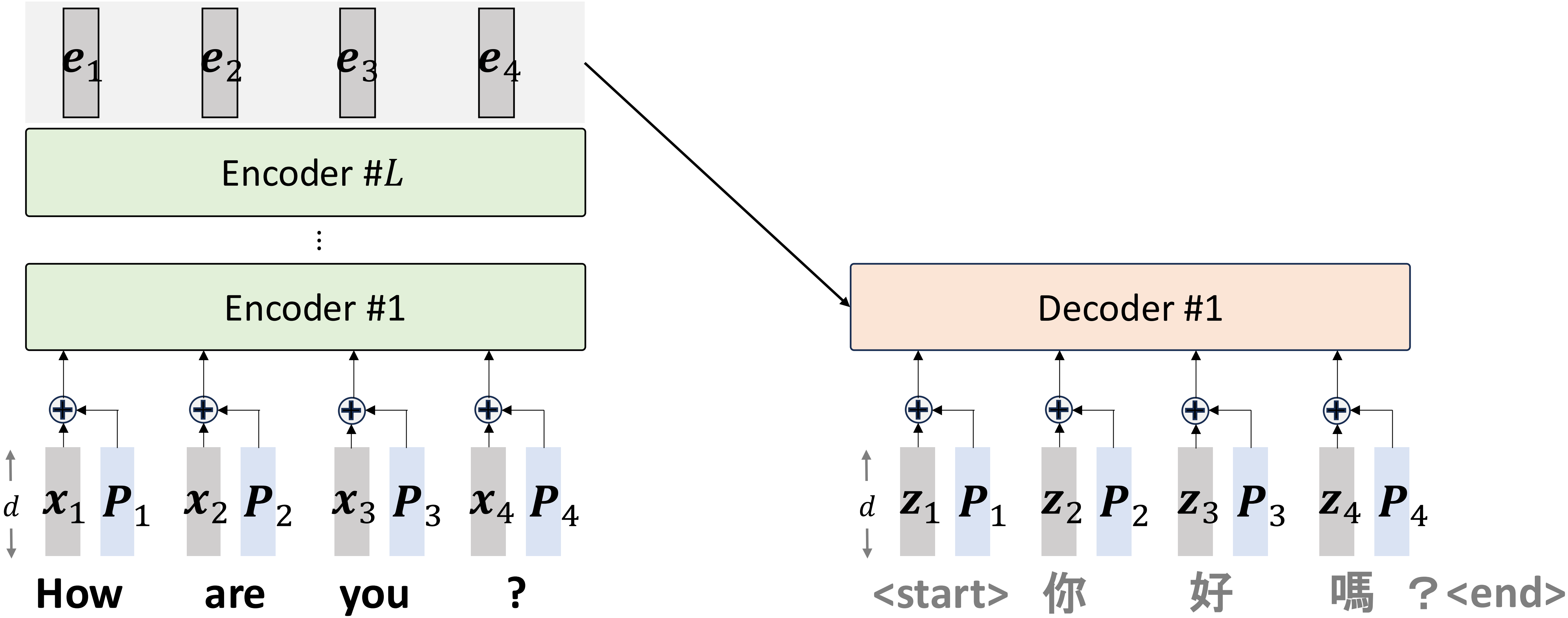
Decoder #1

d x_1 P_1 x_2 P_2 x_3 P_3 x_4 P_4

How are you ?

d z_1 P_1 z_2 P_2 z_3 P_3 z_4 P_4

<start> 你 好 嗎 ？<end>



$\boldsymbol{e}_1$ $\boldsymbol{e}_2$ $\boldsymbol{e}_3$ $\boldsymbol{e}_4$

Encoder #L

⋮

Encoder #1

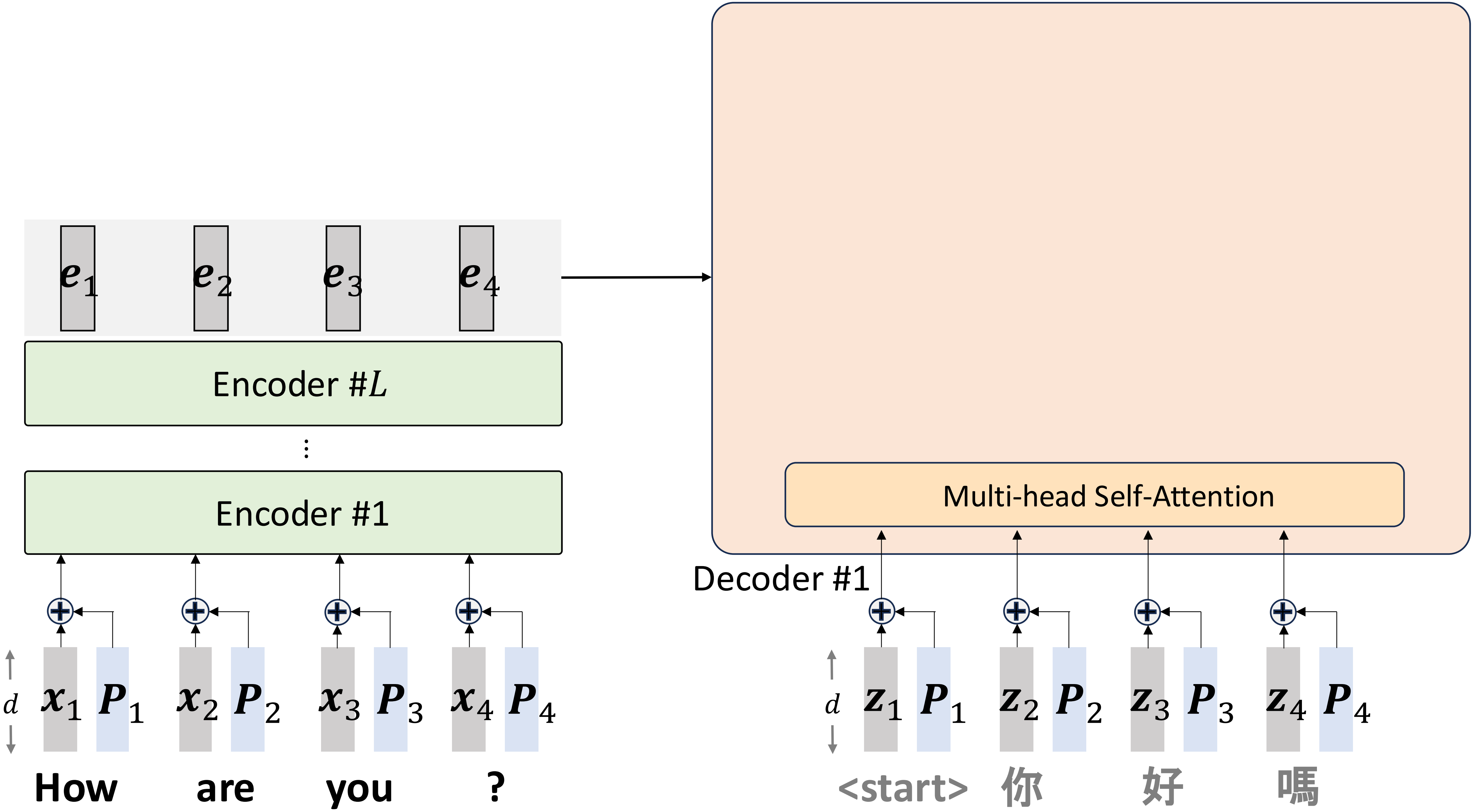
Decoder #1

d $\boldsymbol{x}_1$ $\boldsymbol{P}_1$ $\boldsymbol{x}_2$ $\boldsymbol{P}_2$ $\boldsymbol{x}_3$ $\boldsymbol{P}_3$ $\boldsymbol{x}_4$ $\boldsymbol{P}_4$

How **are** **you** **?**

d $\boldsymbol{z}_1$ $\boldsymbol{P}_1$ $\boldsymbol{z}_2$ $\boldsymbol{P}_2$ $\boldsymbol{z}_3$ $\boldsymbol{P}_3$ $\boldsymbol{z}_4$ $\boldsymbol{P}_4$

<start> **你** **好** **嗎** **？<end>**



e_1 e_2 e_3 e_4

Encoder #L

⋮

Encoder #1

d x_1 P_1 x_2 P_2 x_3 P_3 x_4 P_4

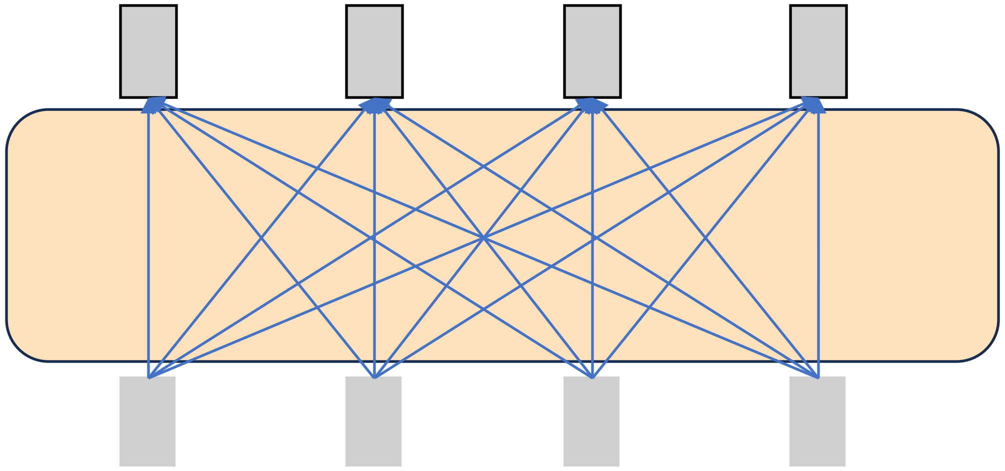
How **are** **you** **?**

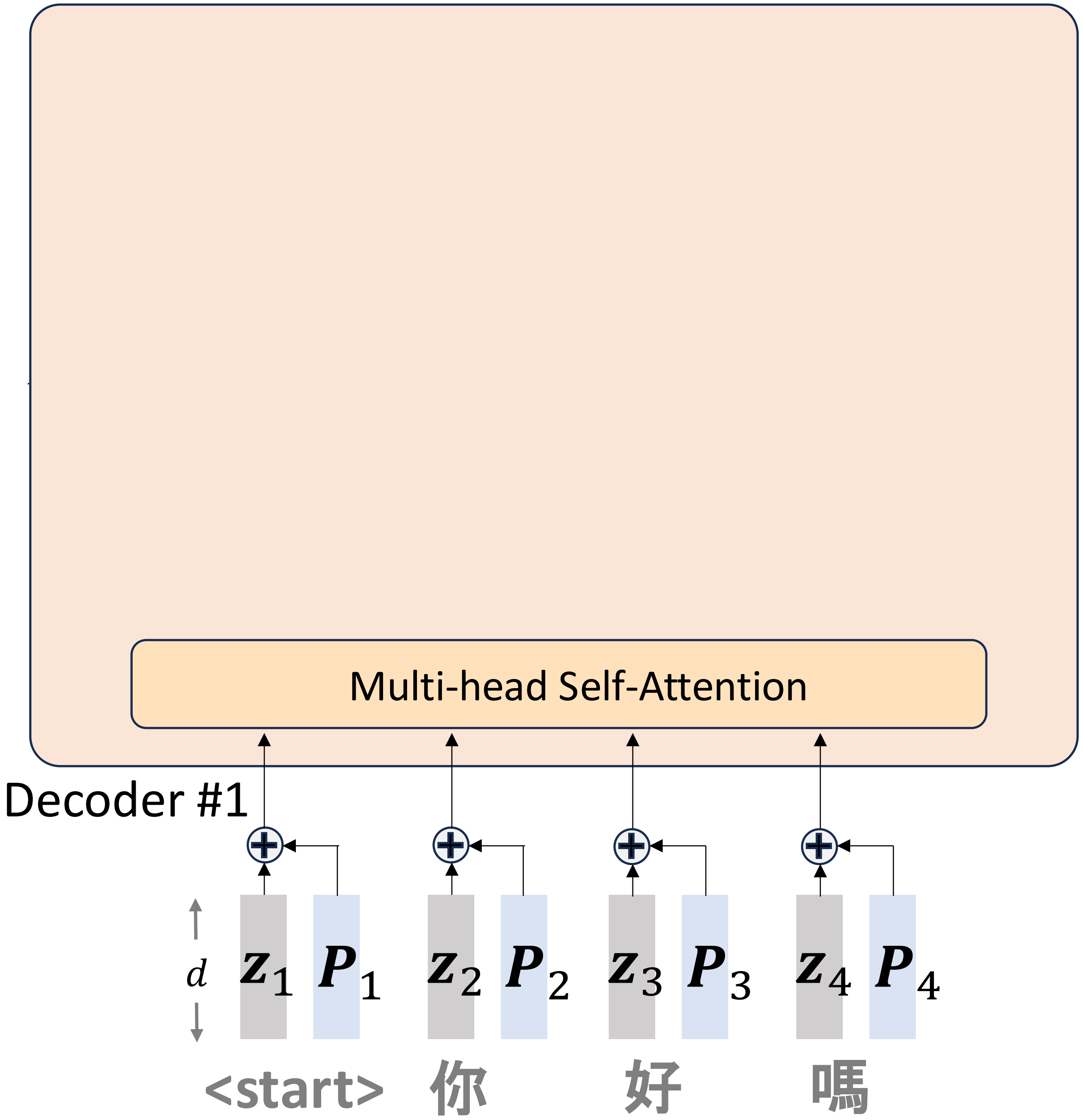
Multi-head Self-Attention

Decoder #1

d z_1 P_1 z_2 P_2 z_3 P_3 z_4 P_4

<start> **你** **好** **嗎**





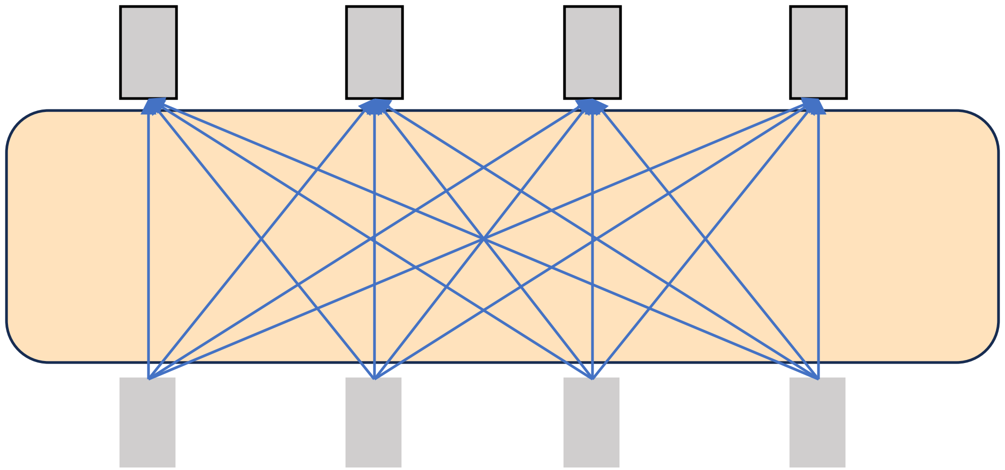
Multi-head Self-Attention

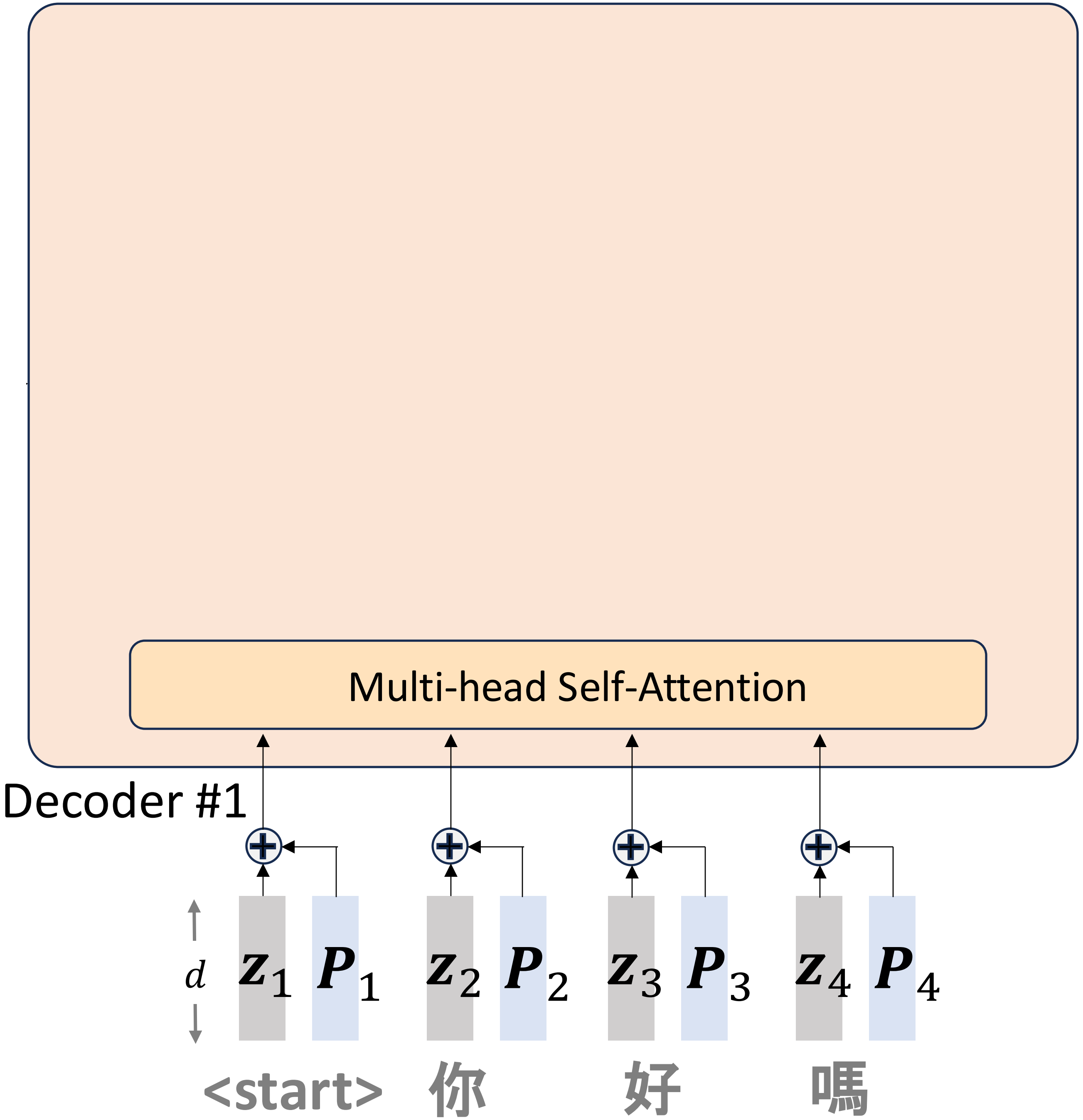
Decoder #1

d

z_1 P_1 z_2 P_2 z_3 P_3 z_4 P_4

<start> 你 好 嗎





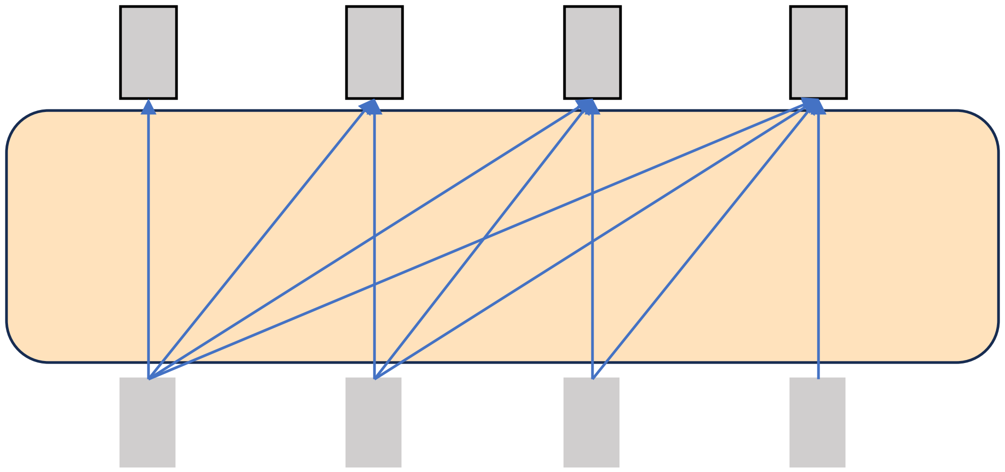
Multi-head Self-Attention

Decoder #1

d

$\boldsymbol{z}_1$ $\boldsymbol{P}_1$ $\boldsymbol{z}_2$ $\boldsymbol{P}_2$ $\boldsymbol{z}_3$ $\boldsymbol{P}_3$ $\boldsymbol{z}_4$ $\boldsymbol{P}_4$

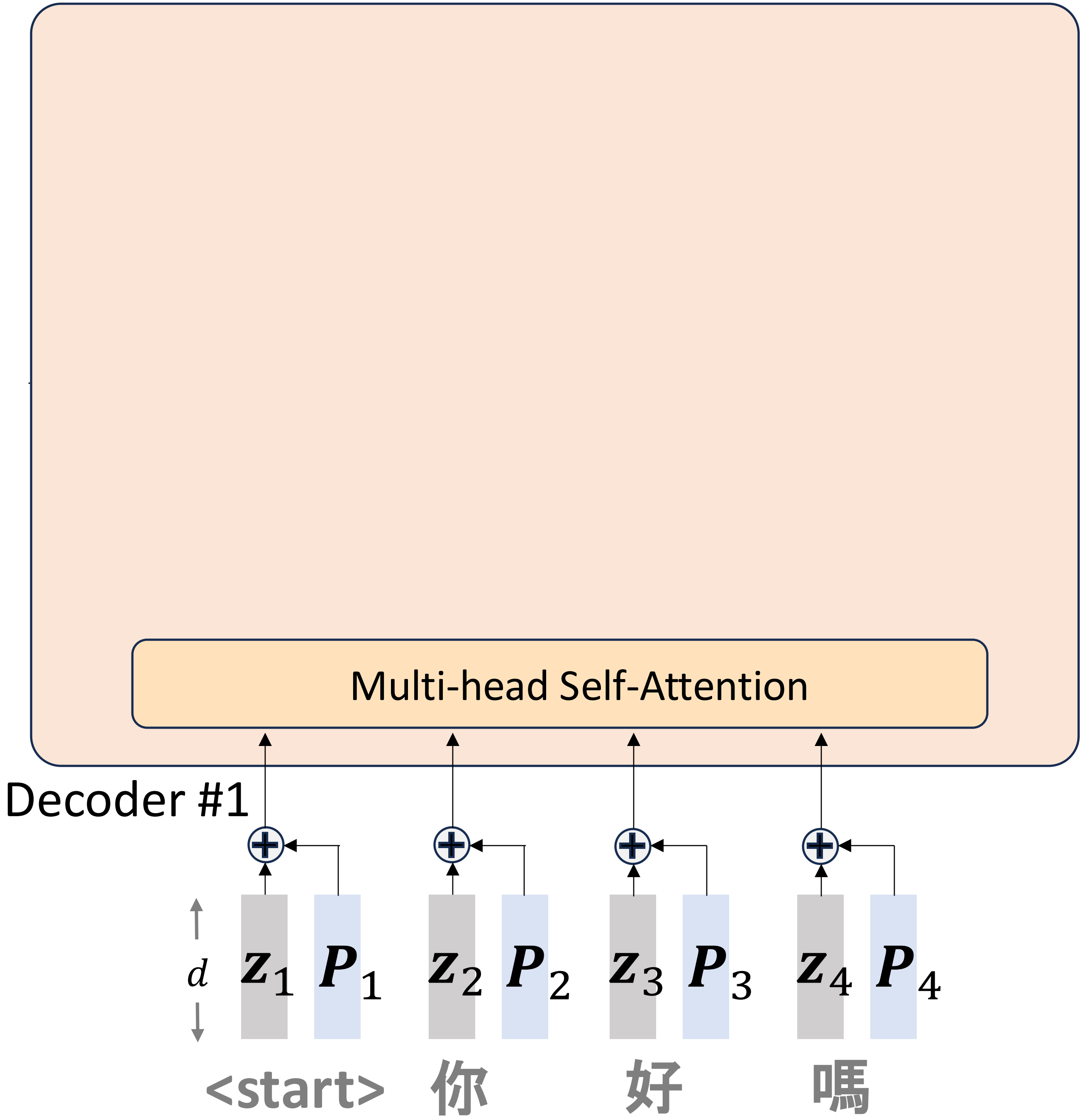
<start> 你 好 嗎

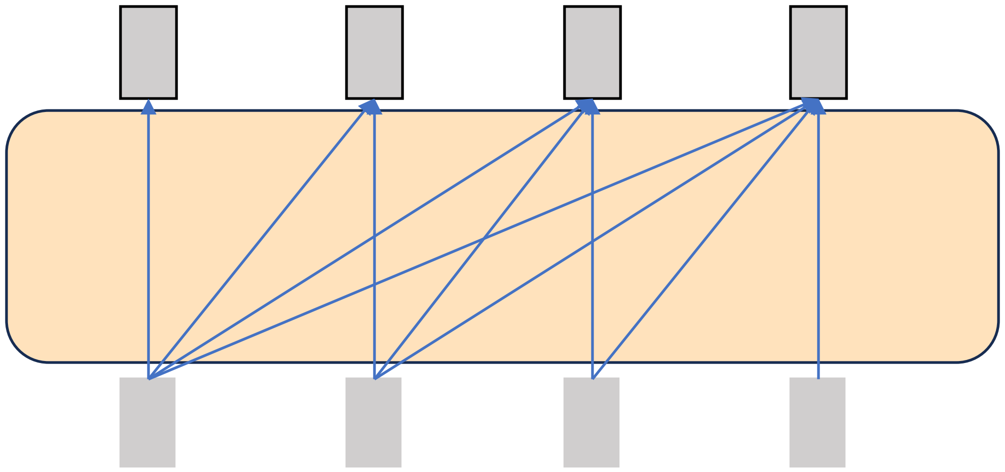


$$\text{Attention}(Q,K,V) = V\ \text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$

$$\text{MaskedAttention}(Q,K,V) = V\ \text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}} + M\right)$$

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ -\infty & 0 & 0 & 0 & 0 \\ -\infty & -\infty & 0 & 0 & 0 \\ -\infty & -\infty & -\infty & 0 & 0 \\ -\infty & -\infty & -\infty & -\infty & 0 \end{bmatrix}$$

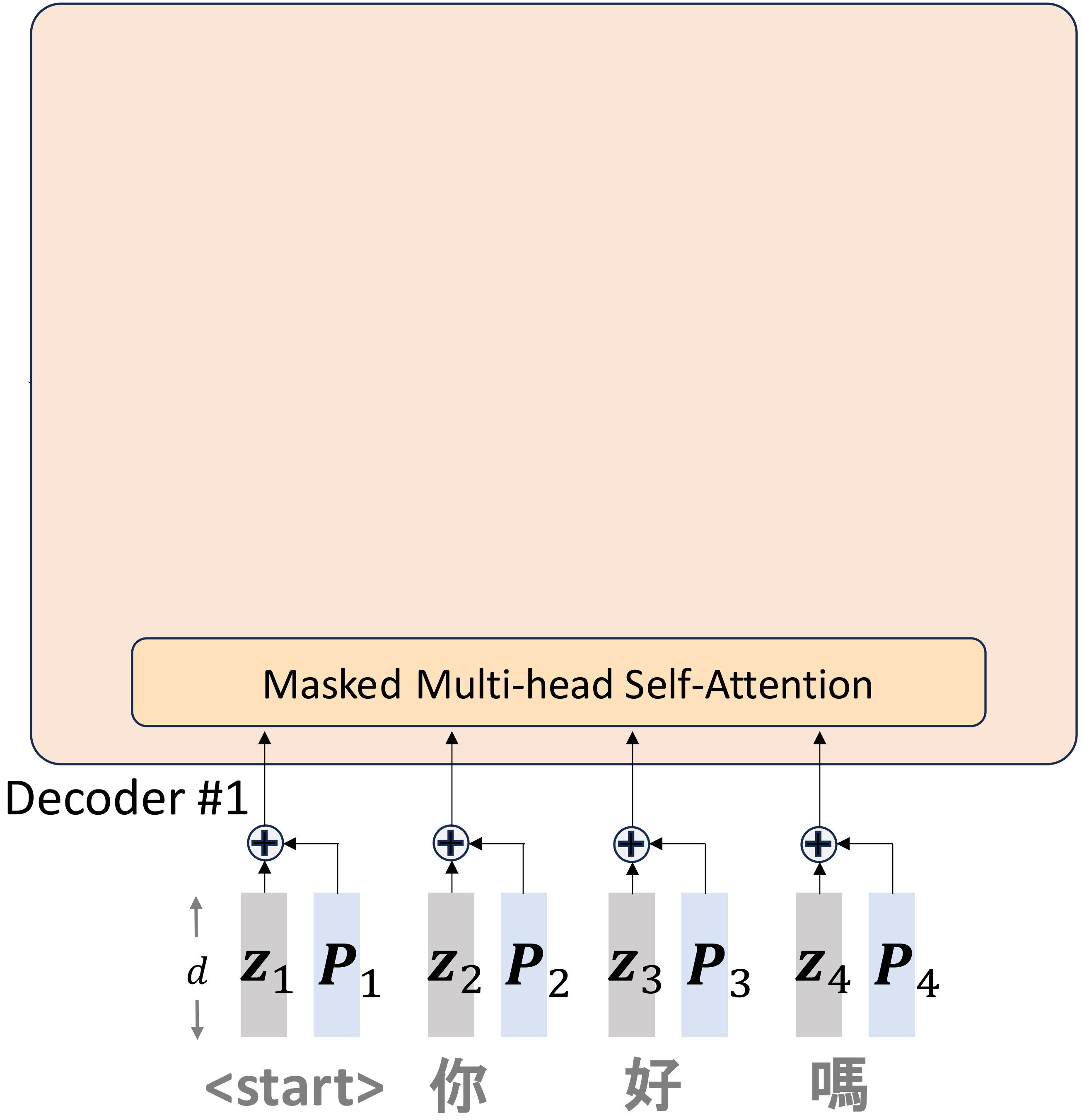


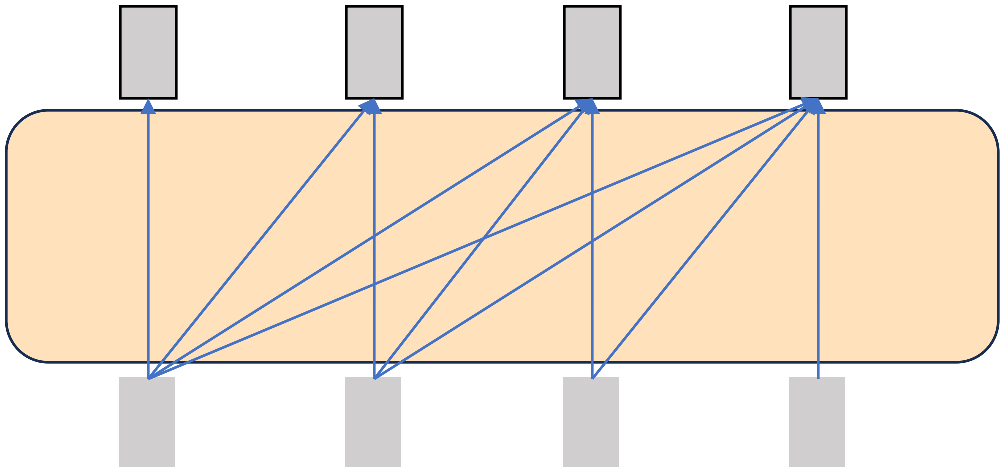


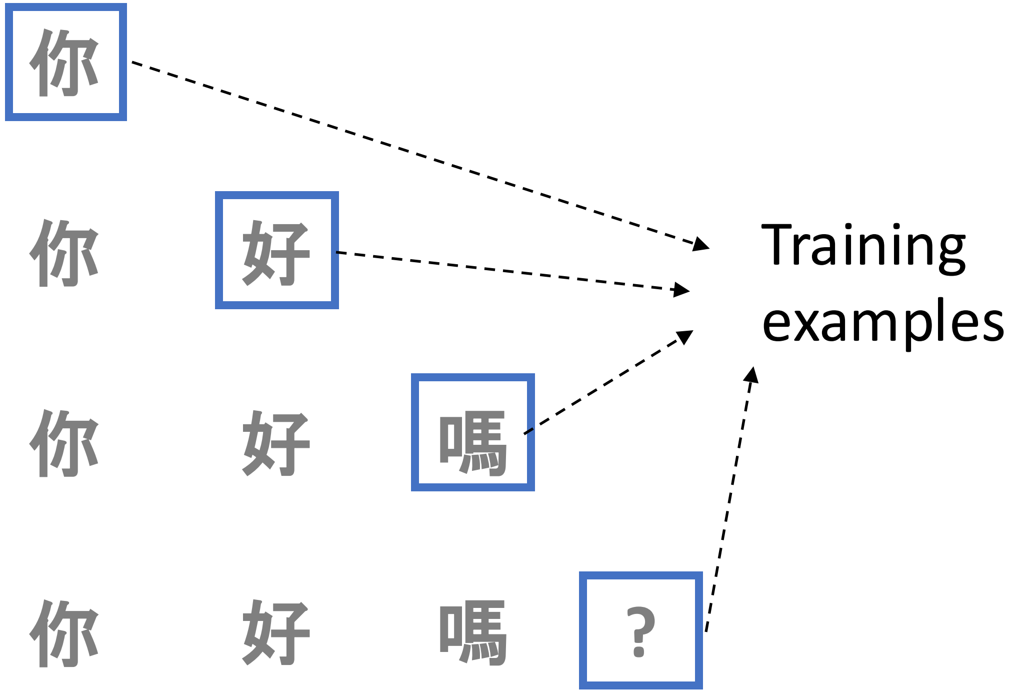
$$\text{Attention}(Q,K,V) = V\,\text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}}\right)$$

$$\text{MaskedAttention}(Q,K,V) = V\,\text{softmax}\left(\frac{K^\top Q}{\sqrt{d_k}} + M\right)$$

$$M = \begin{array}{|c|c|c|c|c|} \hline 0 & 0 & 0 & 0 & 0 \\ \hline -\infty & 0 & 0 & 0 & 0 \\ \hline -\infty & -\infty & 0 & 0 & 0 \\ \hline -\infty & -\infty & -\infty & 0 & 0 \\ \hline -\infty & -\infty & -\infty & -\infty & 0 \\ \hline \end{array}$$







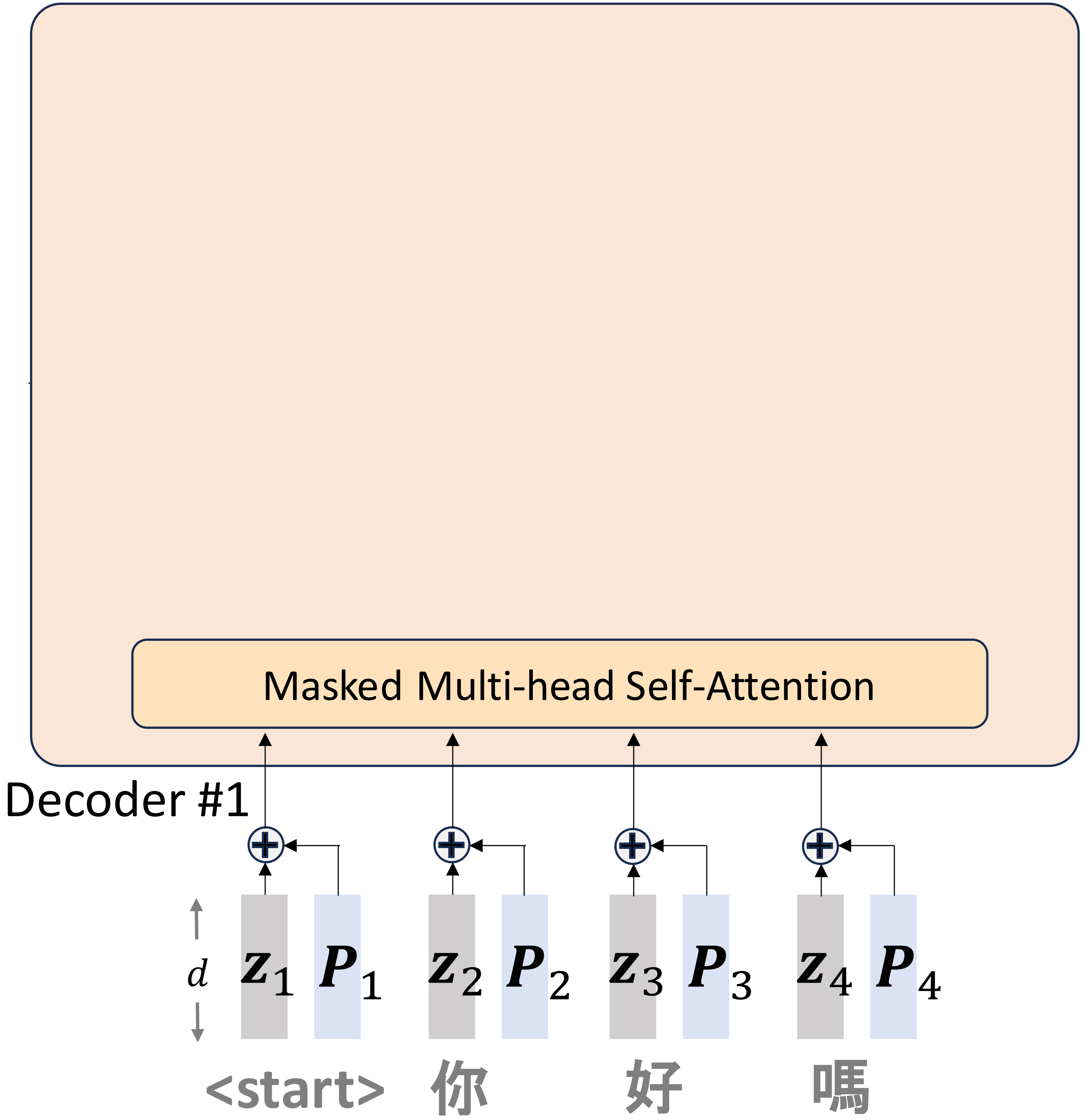
你

你 好

你 好 嗎

你 好 嗎 ?

Training examples



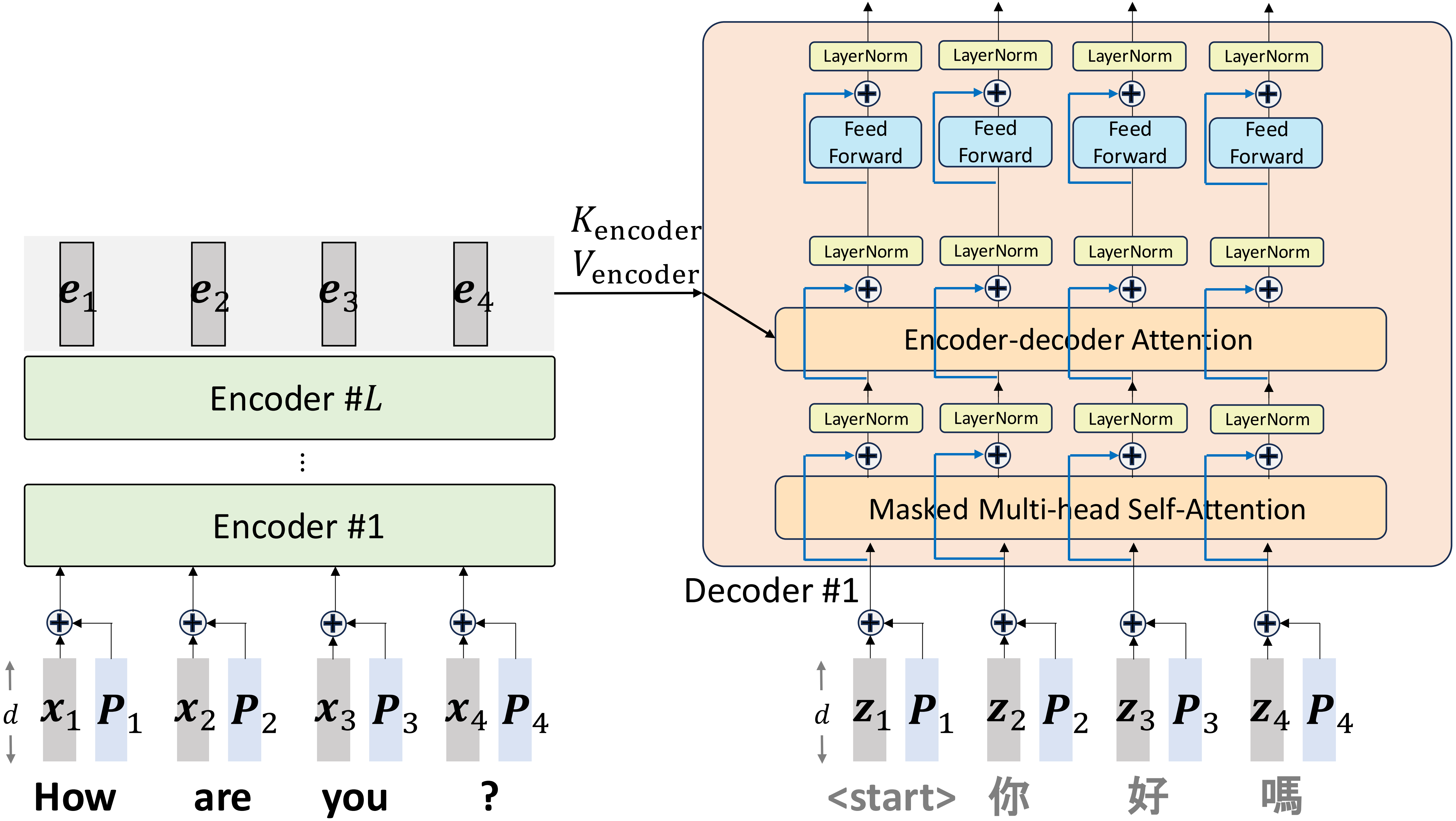
Masked Multi-head Self-Attention

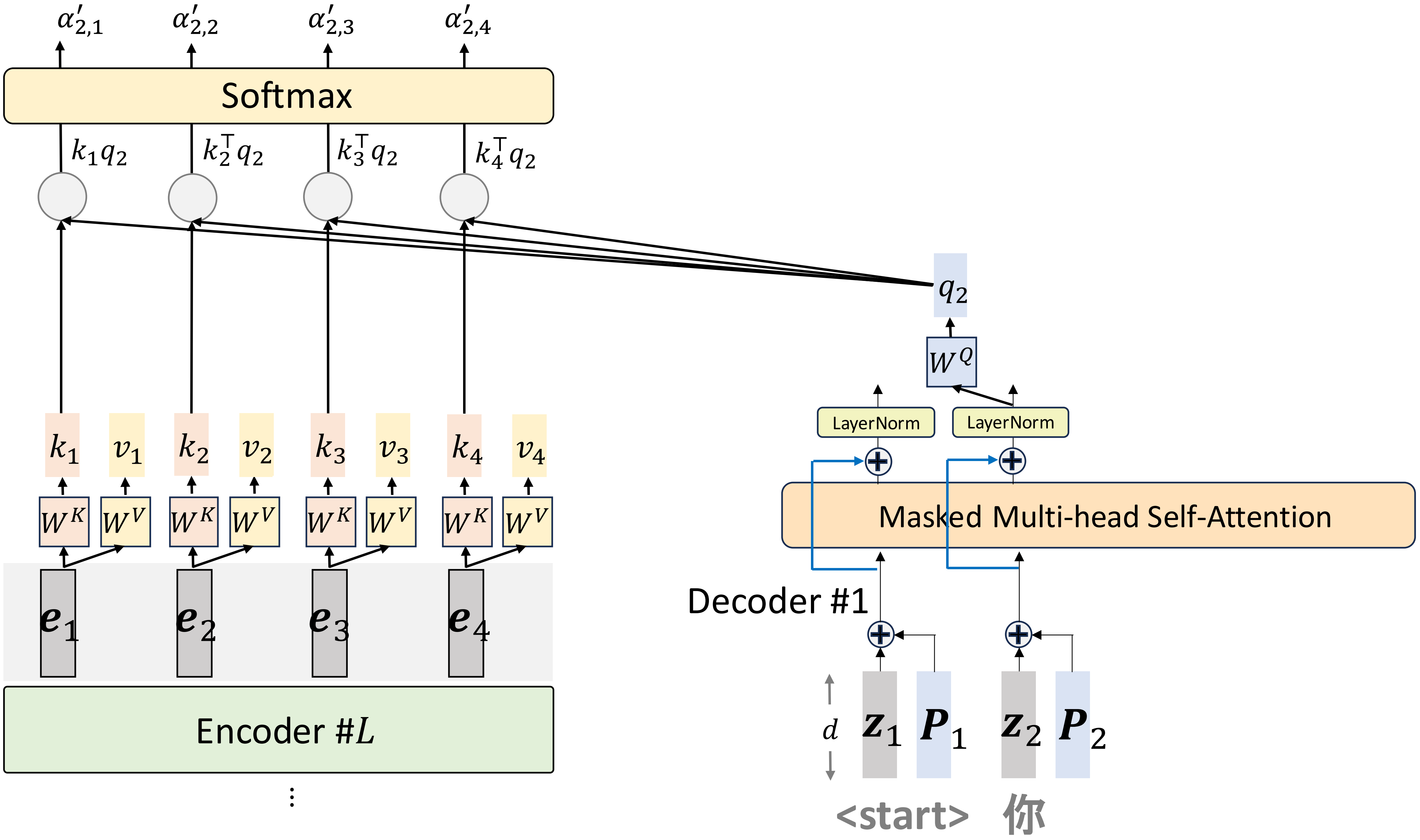
Decoder #1

d

$\boldsymbol{z}_1$ $\boldsymbol{P}_1$ $\boldsymbol{z}_2$ $\boldsymbol{P}_2$ $\boldsymbol{z}_3$ $\boldsymbol{P}_3$ $\boldsymbol{z}_4$ $\boldsymbol{P}_4$

<start> 你 好 嗎


LayerNorm
Feed Forward
LayerNorm
Encoder-decoder Attention
LayerNorm
Masked Multi-head Self-Attention
Decoder #1
K_{encoder}
V_{encoder}
e_1
e_2
e_3
e_4
Encoder #L
⋮
Encoder #1
d
x_1 P_1
x_2 P_2
x_3 P_3
x_4 P_4
How
are
you
?
d
z_1 P_1
z_2 P_2
z_3 P_3
z_4 P_4
<start>
你
好
嗎

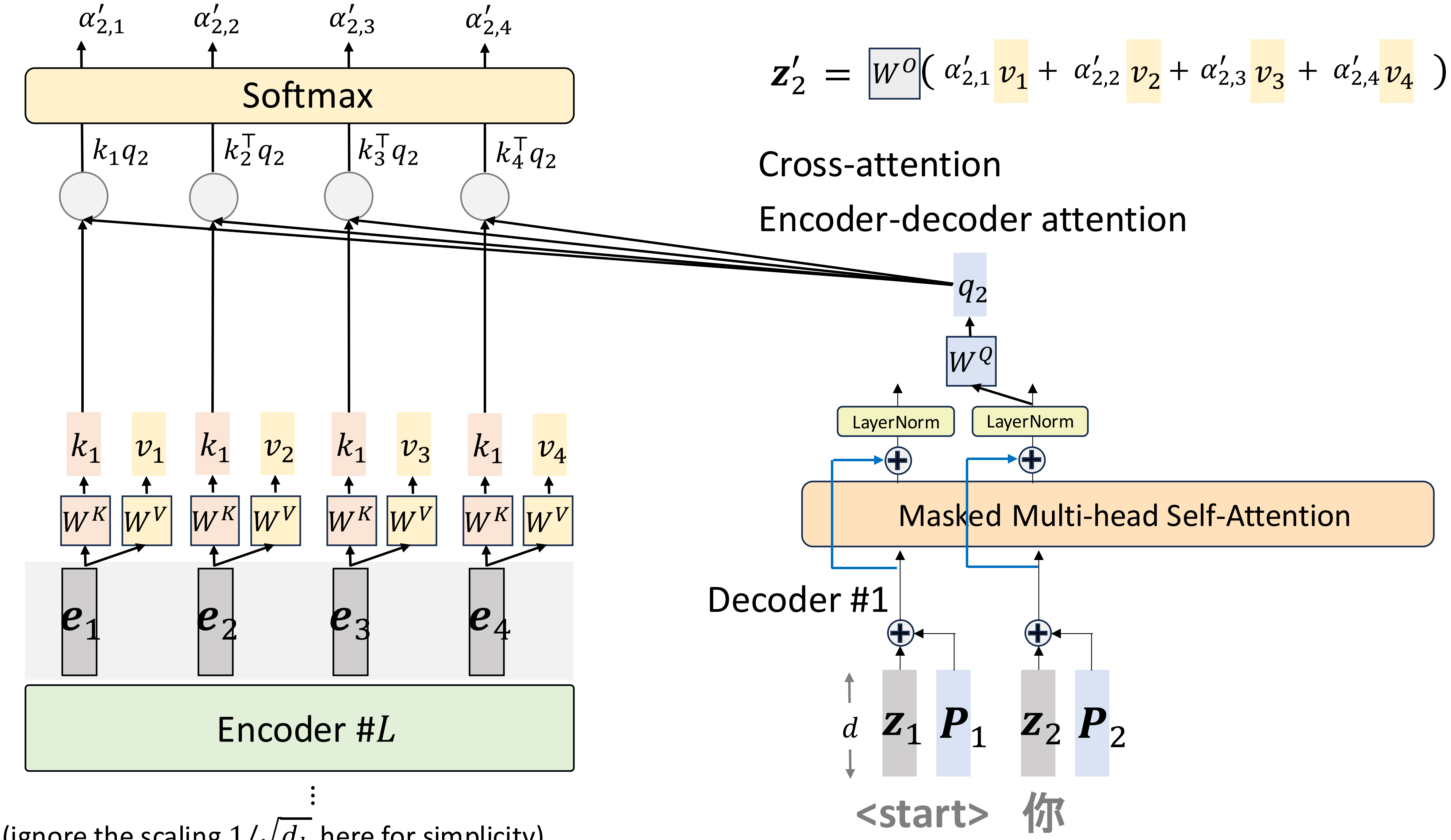


$\alpha'_{2,1}$
$\alpha'_{2,2}$
$\alpha'_{2,3}$
$\alpha'_{2,4}$
Softmax
$k_1 q_2$
$k_2^\top q_2$
$k_3^\top q_2$
$k_4^\top q_2$
q_2
W^Q
LayerNorm
LayerNorm
Masked Multi-head Self-Attention
Decoder #1
k_1
v_1
k_2
v_2
k_3
v_3
k_4
v_4
W^K
W^V
W^K
W^V
W^K
W^V
W^K
W^V
$\boldsymbol{e}_1$
$\boldsymbol{e}_2$
$\boldsymbol{e}_3$
$\boldsymbol{e}_4$
Encoder #L
⋮
d
$\boldsymbol{z}_1$
$\boldsymbol{P}_1$
$\boldsymbol{z}_2$
$\boldsymbol{P}_2$
<start>
你

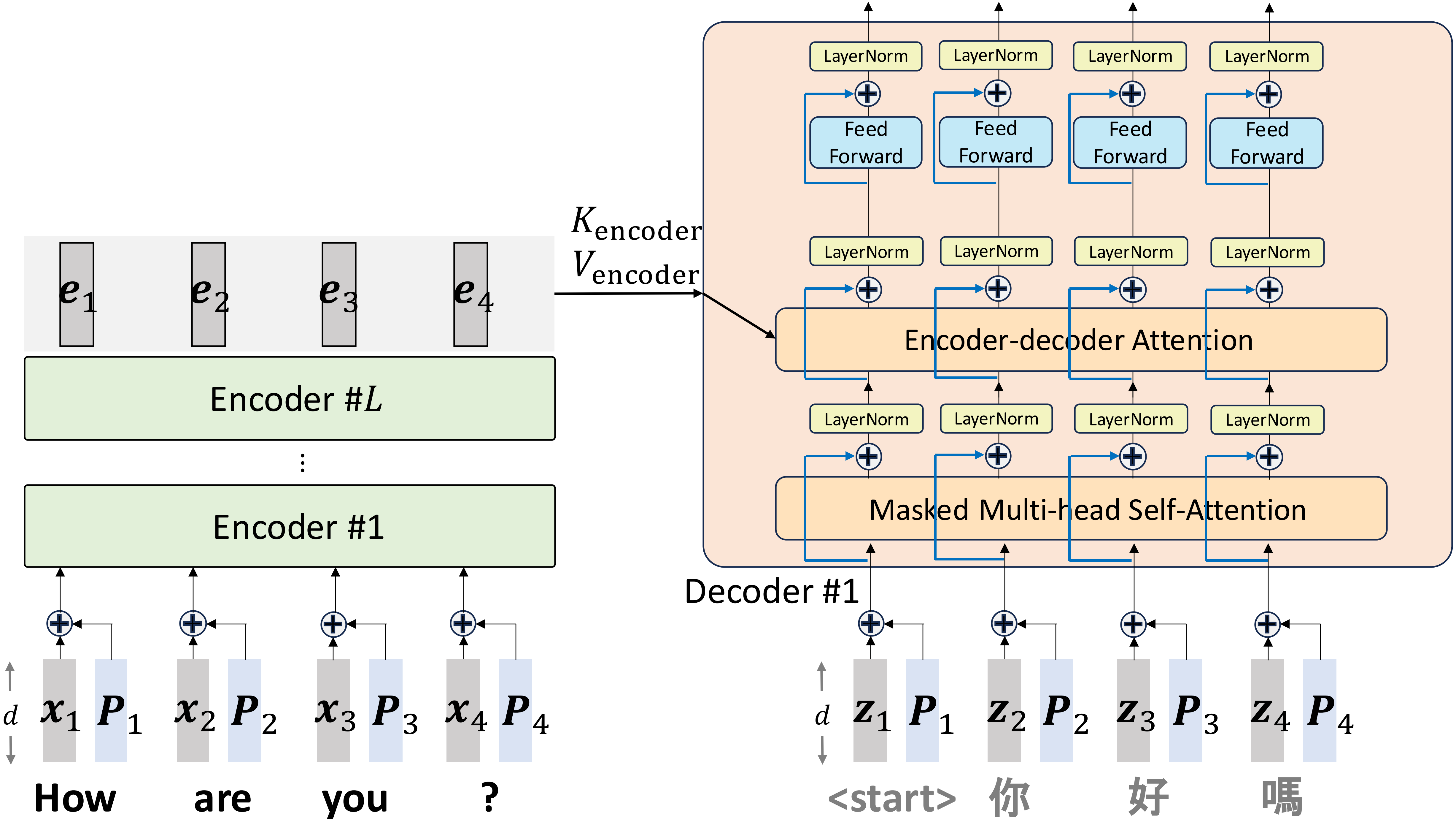
$$\boldsymbol{z}'_2 = W^O(\alpha'_{2,1} v_1 + \alpha'_{2,2} v_2 + \alpha'_{2,3} v_3 + \alpha'_{2,4} v_4)$$

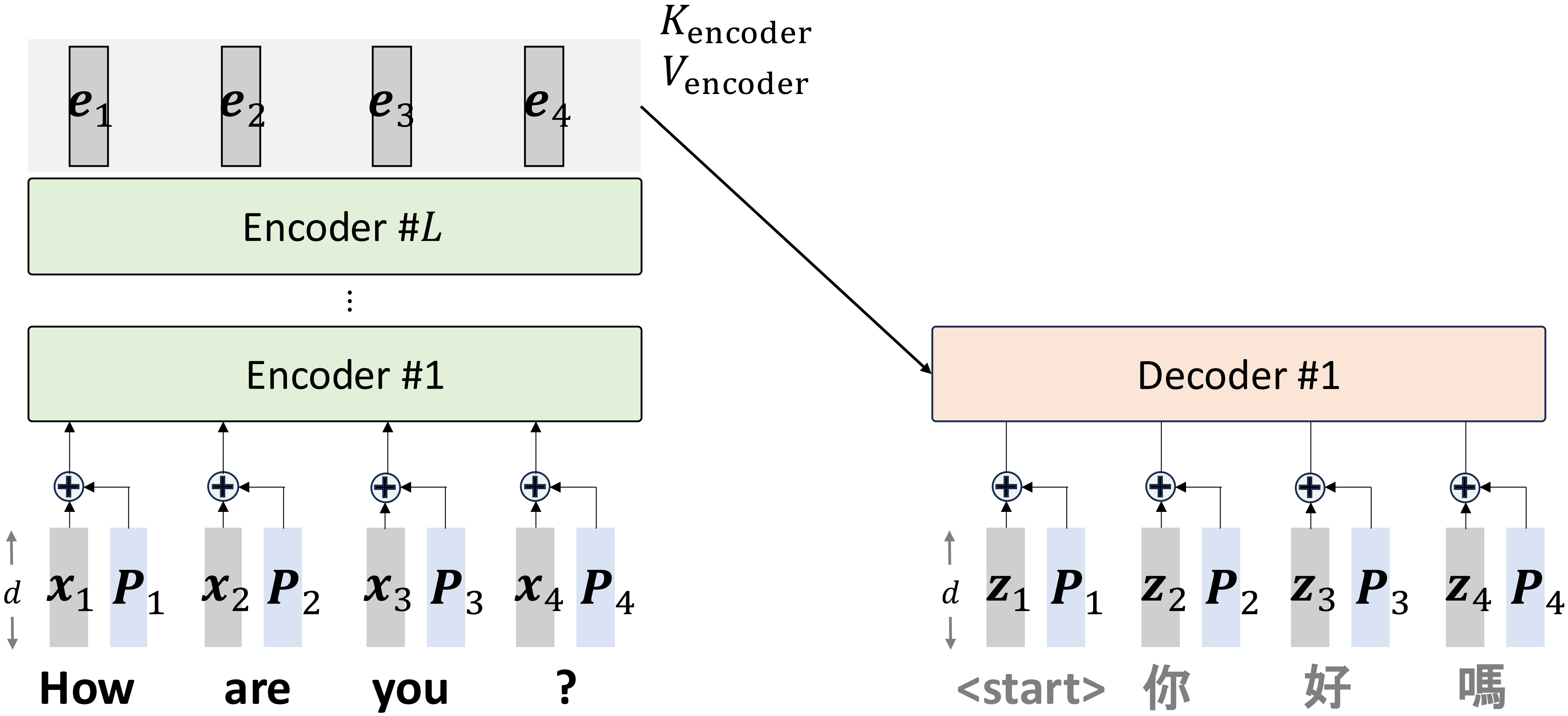
Cross-attention

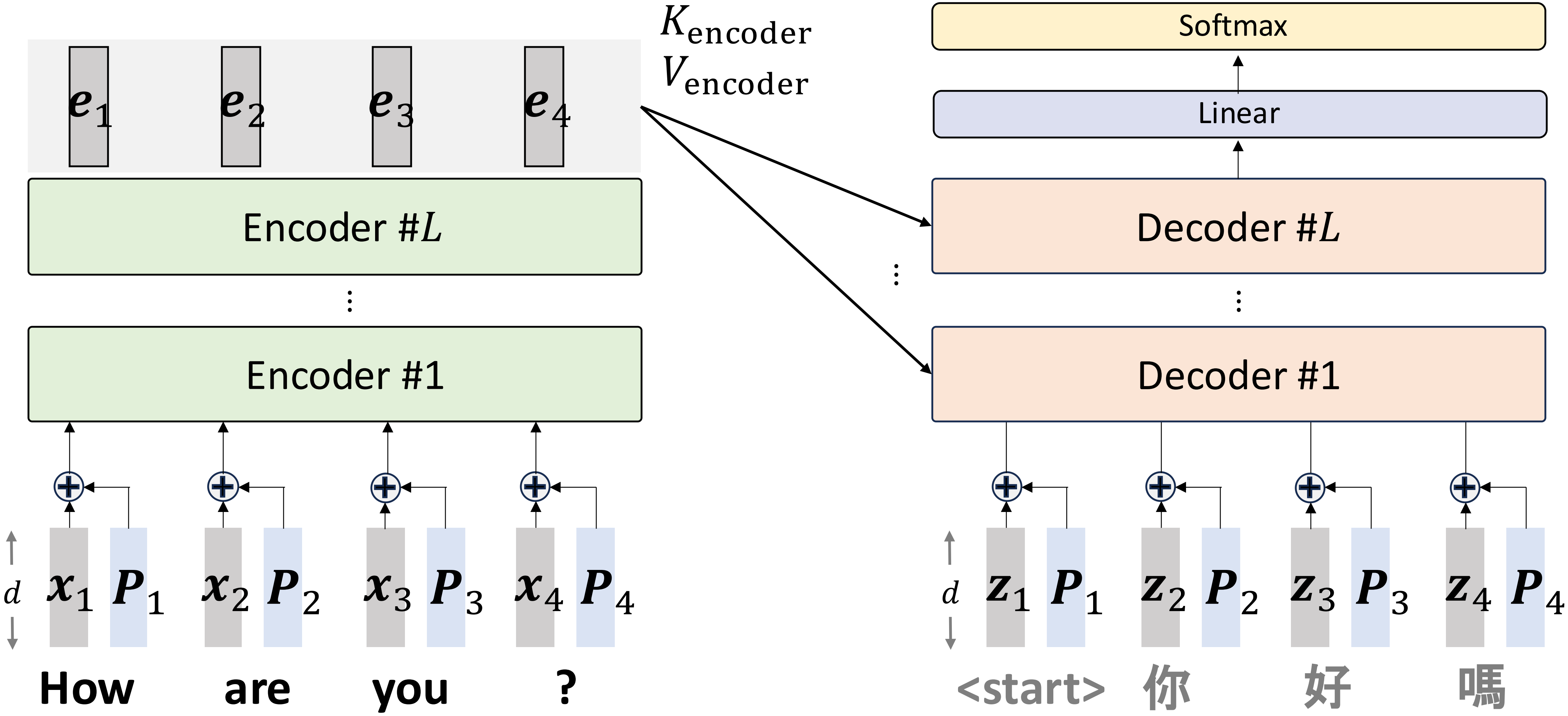
Encoder-decoder attention



(ignore the scaling $1/\sqrt{d_k}$ here for simplicity)


LayerNorm
LayerNorm
LayerNorm
LayerNorm
Feed Forward
Feed Forward
Feed Forward
Feed Forward
K_{encoder}
V_{encoder}
LayerNorm
LayerNorm
LayerNorm
LayerNorm
e_1
e_2
e_3
e_4
Encoder-decoder Attention
Encoder #L
LayerNorm
LayerNorm
LayerNorm
LayerNorm
⋮
Masked Multi-head Self-Attention
Encoder #1
Decoder #1
d
x_1 P_1
x_2 P_2
x_3 P_3
x_4 P_4
d
z_1 P_1
z_2 P_2
z_3 P_3
z_4 P_4
How
are
you
?
<start>
你
好
嗎





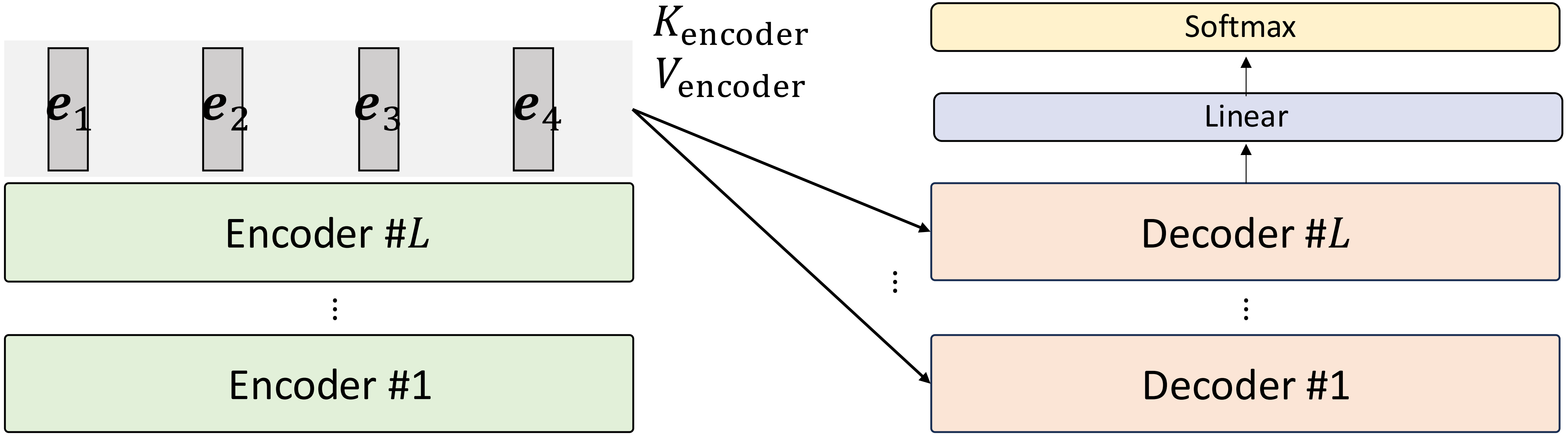
Encoder-Decoder Transformer

Examples:

Attention is all you need, T5, BART.

Good for:

Machine translation, summarization. QA (when input/target are sufficiently different)



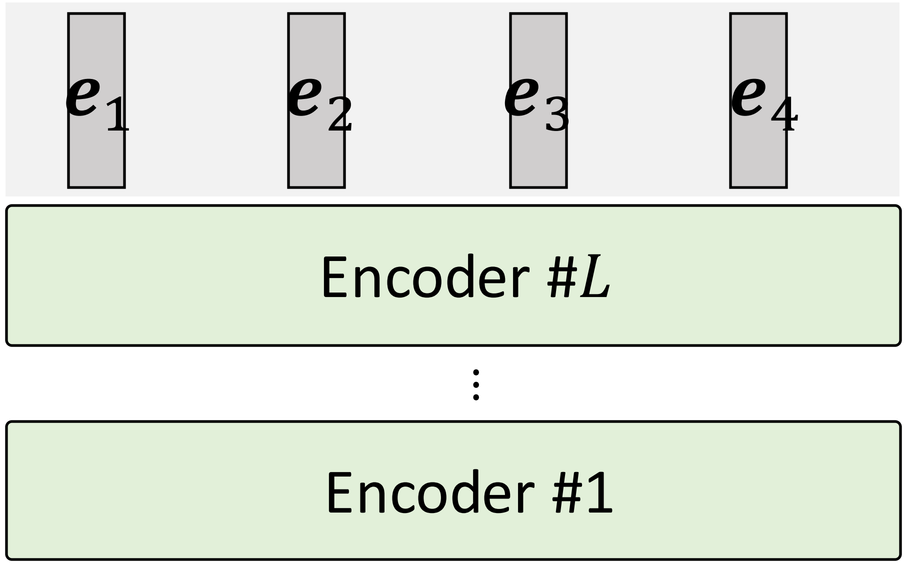
Encoder-Decoder Transformer

Examples:

Attention is all you need, T5, BART.

Good for:

Machine translation, summarization. QA (when input/target are sufficiently different)



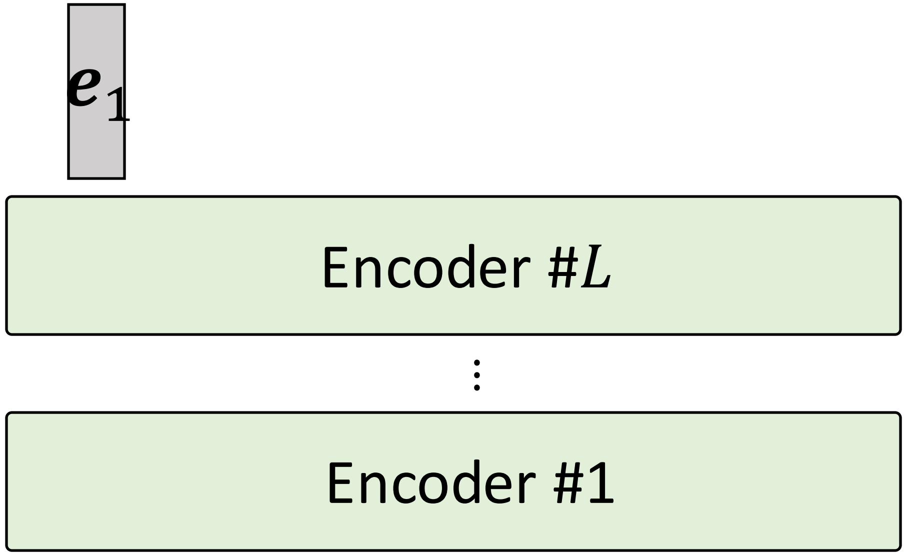
Encoder-only Transformer

Examples:

BERT, RoBERTa, DeBERTa, X-BERT

Good for:

Classification, sequence tagging, sentiment analysis
(Understand text, but not generate them)



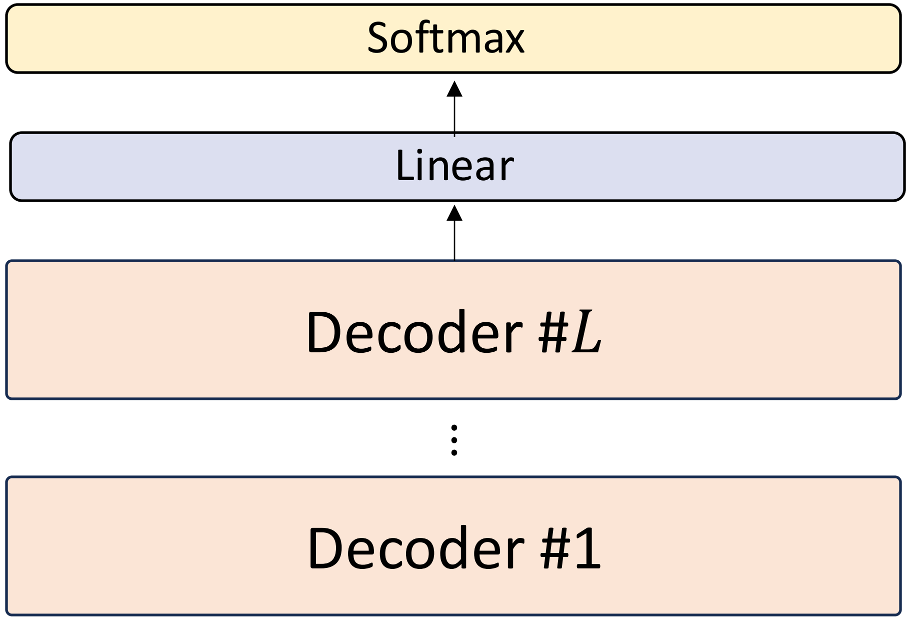
Decoder-only Transformer

Examples:

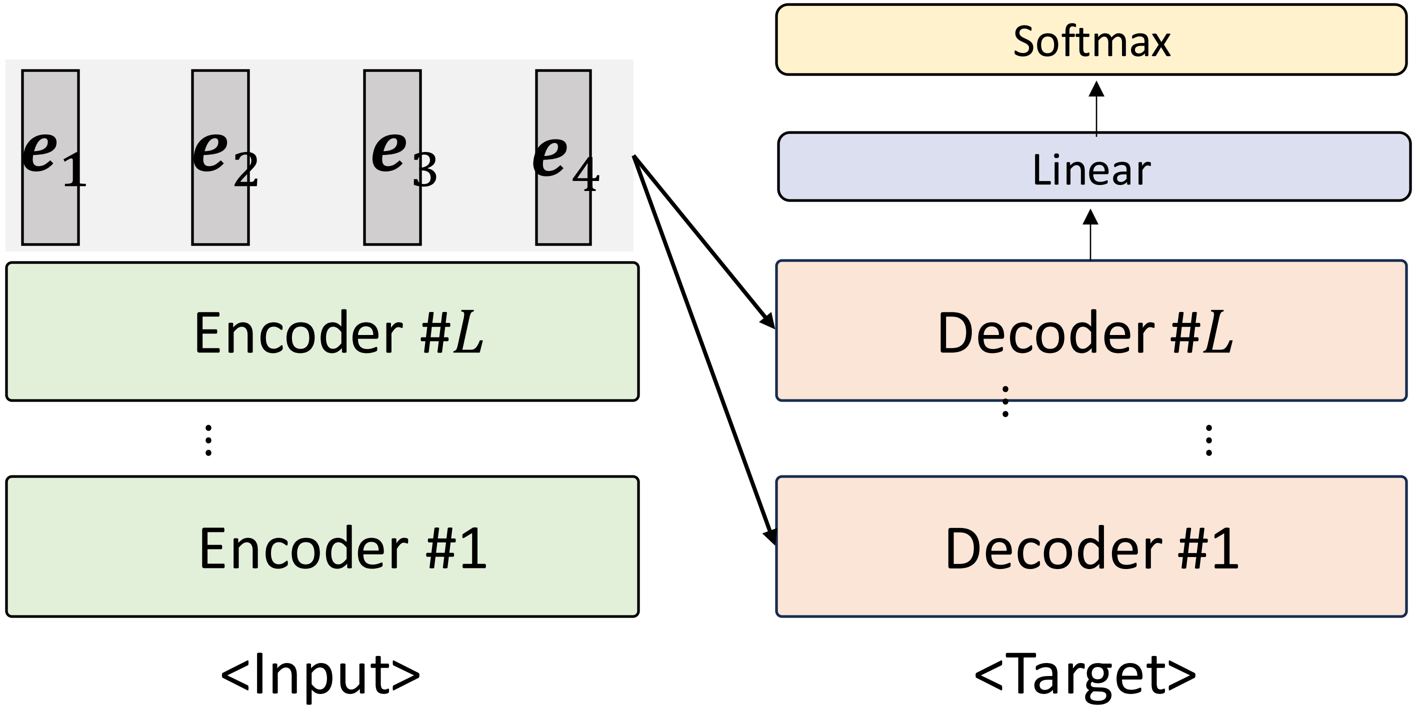
GPT-X (OpenAI), PaLM (Google), LLaMA (Meta)
BLOOM (BigScience)

Good for:

Text generation, multi-round conversation

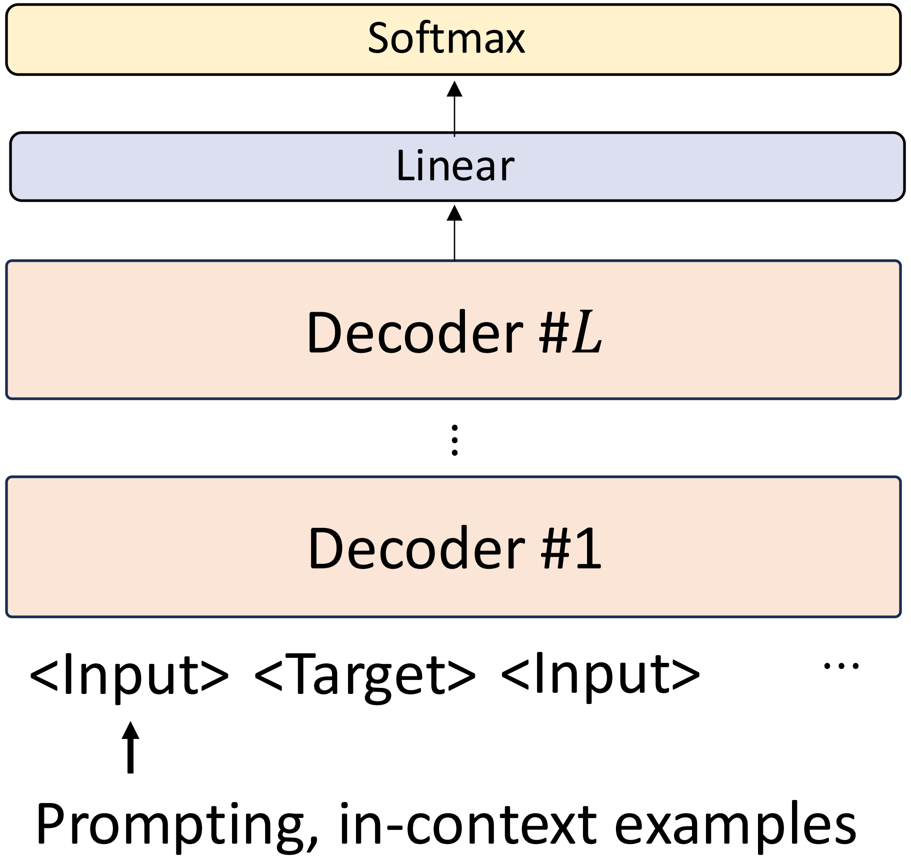


Encoder-Decoder Transformer



Different parameters for encoder/decoder

Decoder-only Transformer



Shared parameters

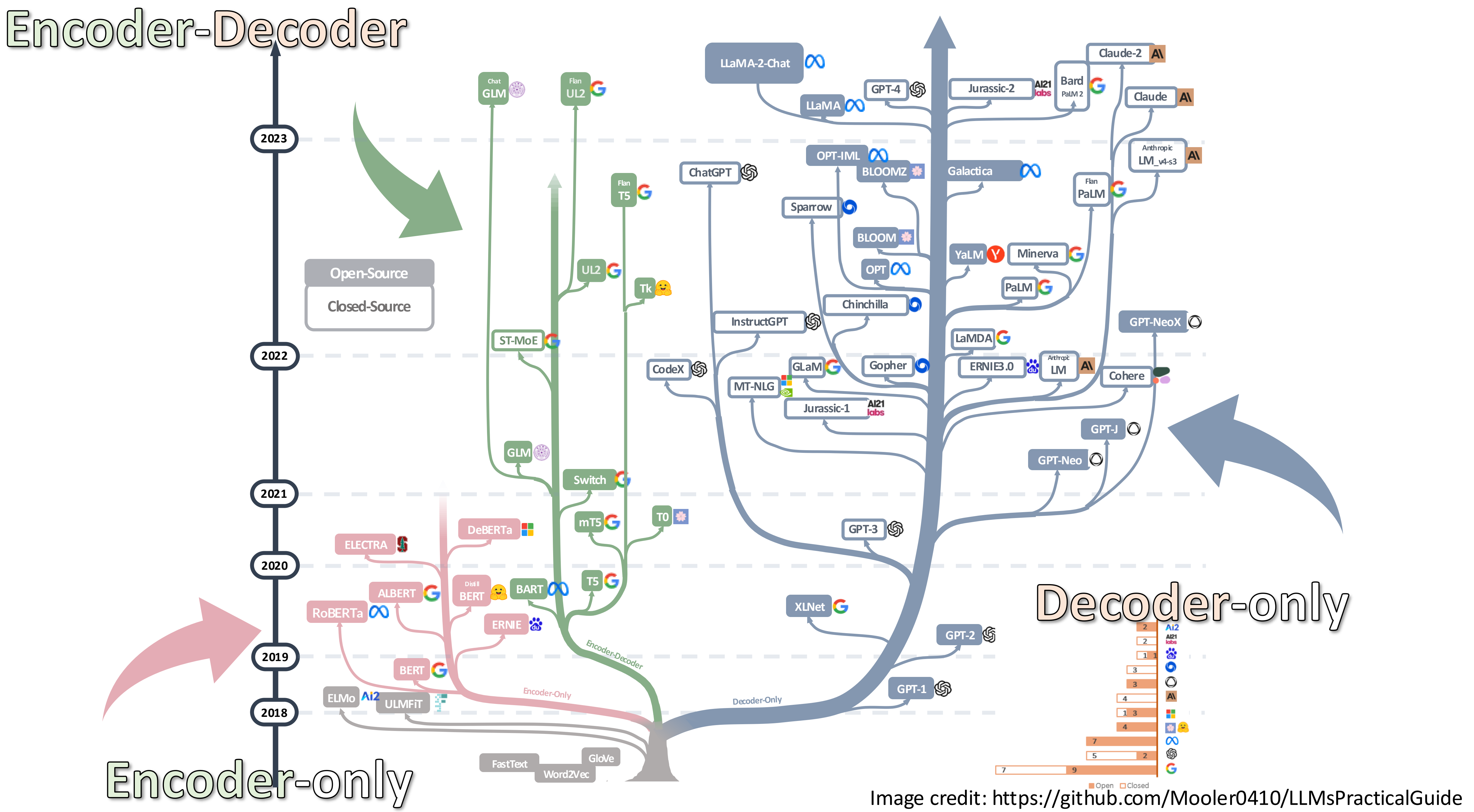
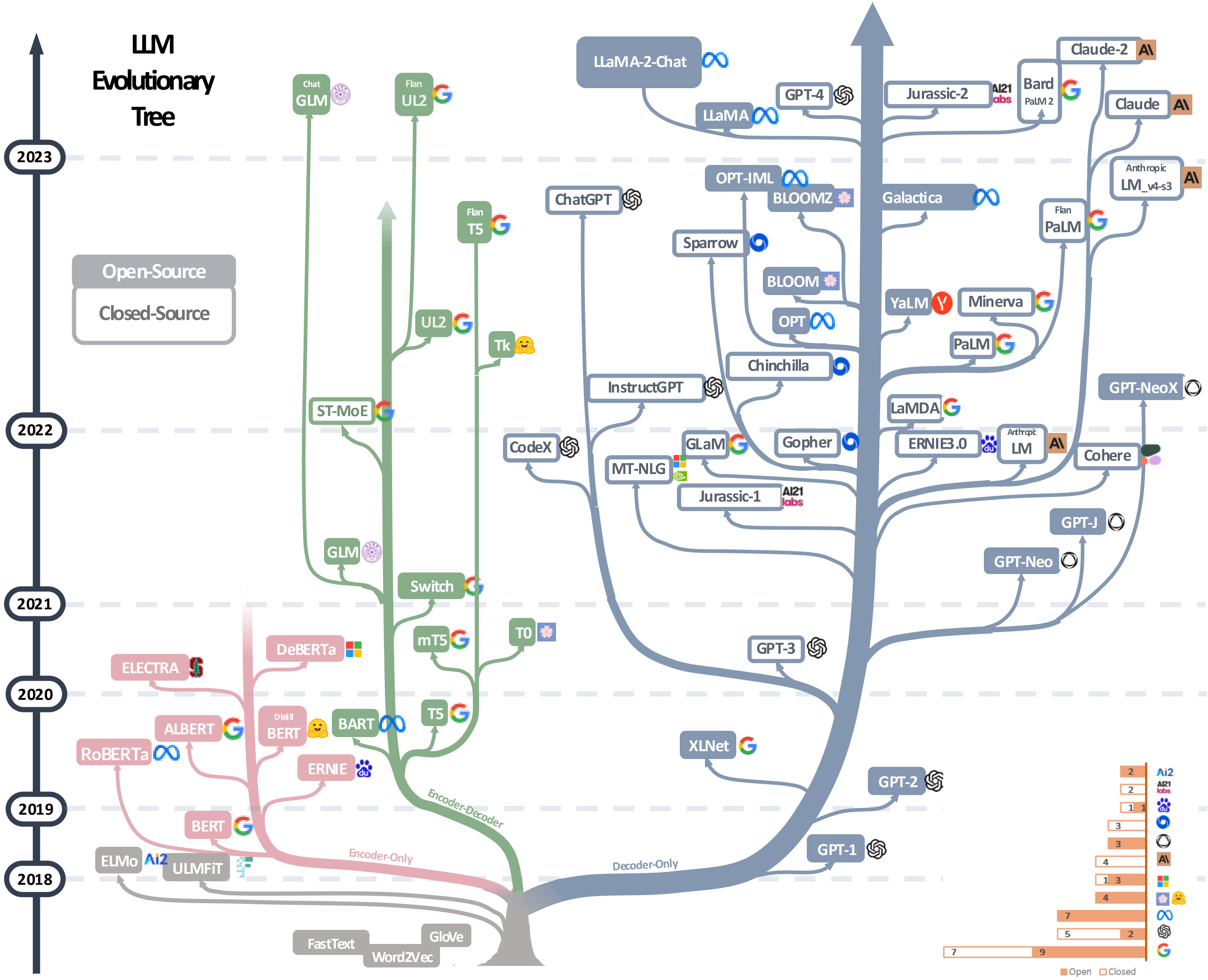


Image credit: https://github.com/Mooler0410/LLMsPracticalGuide


LLM Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Encoder-Only
Encoder-Decoder
Decoder-Only
FastText
Word2Vec
GloVe
ELMo
ULMFiT
BERT
RoBERTa
ALBERT
Distill BERT
ERNIE
ELECTRA
DeBERTa
BART
T5
mT5
T0
Switch
GLM
ST-MoE
UL2
Tk
Flan T5
Flan UL2
Chat GLM
GPT-1
GPT-2
XLNet
GPT-3
GPT-Neo
GPT-J
GPT-NeoX
Cohere
CodeX
MT-NLG
GLaM
Jurassic-1
Gopher
ERNIE3.0
Anthropic LM
LaMDA
InstructGPT
Chinchilla
OPT
BLOOM
YaLM
Minerva
PaLM
Sparrow
BLOOMZ
OPT-IML
Galactica
Flan PaLM
Anthropic LM_v4-s3
ChatGPT
LLaMA
GPT-4
LLaMA-2-Chat
Jurassic-2
Bard PaLM 2
Claude
Claude-2
2
2
1 1
3
3
4
1 3
4
7
5 2
7 9
Open
Closed