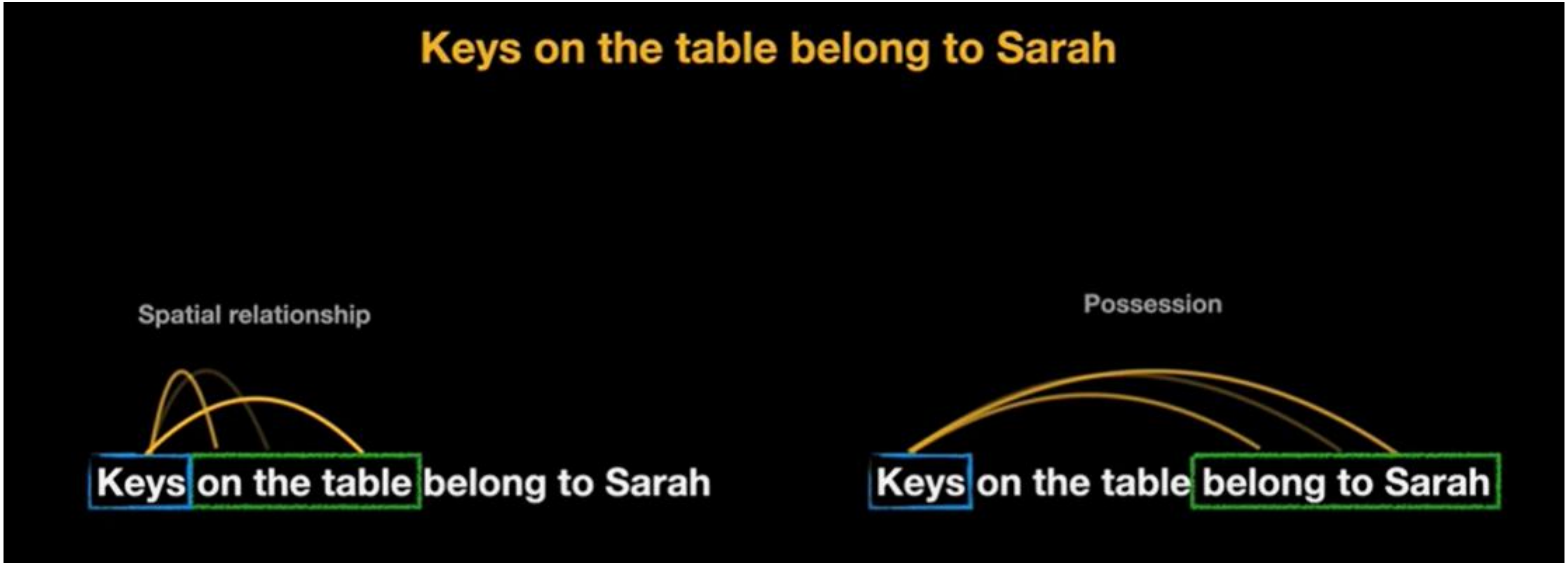
Deep Learning : Multi-headed Attention in Transformers

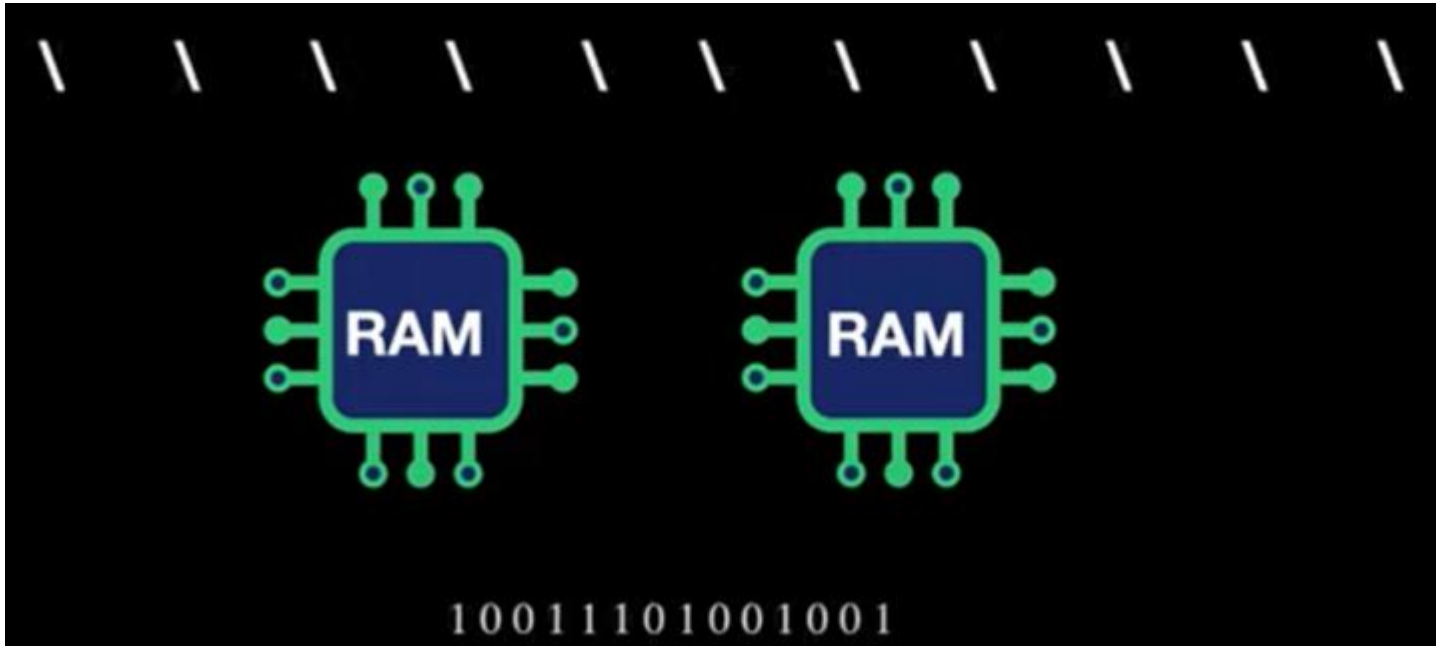
राष्ट्रीय प्रौद्योगिकी संस्थान सिक्किम
NATIONAL INSTITUTE OF TECHNOLOGY SIKKIM

Course Instructor:
Dr. Bam Bahadur Sinha
Assistant Professor
Computer Science & Engineering
National Institute of Technology
Sikkim



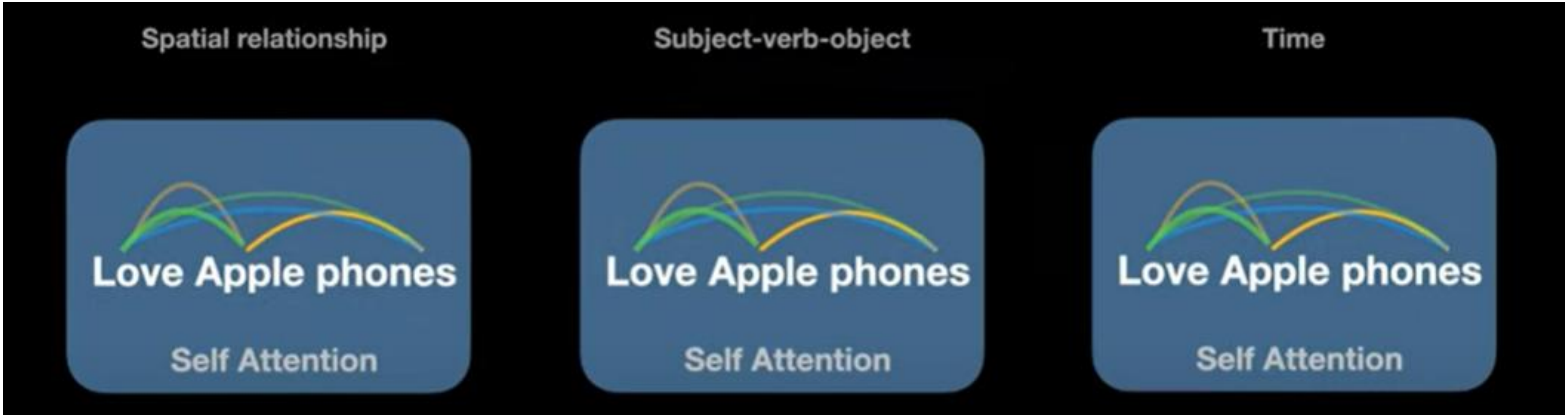
Keys on the table belong to Sarah
Spatial relationship
Keys on the table belong to Sarah
Possession
Keys on the table belong to Sarah

Multi-headed attention

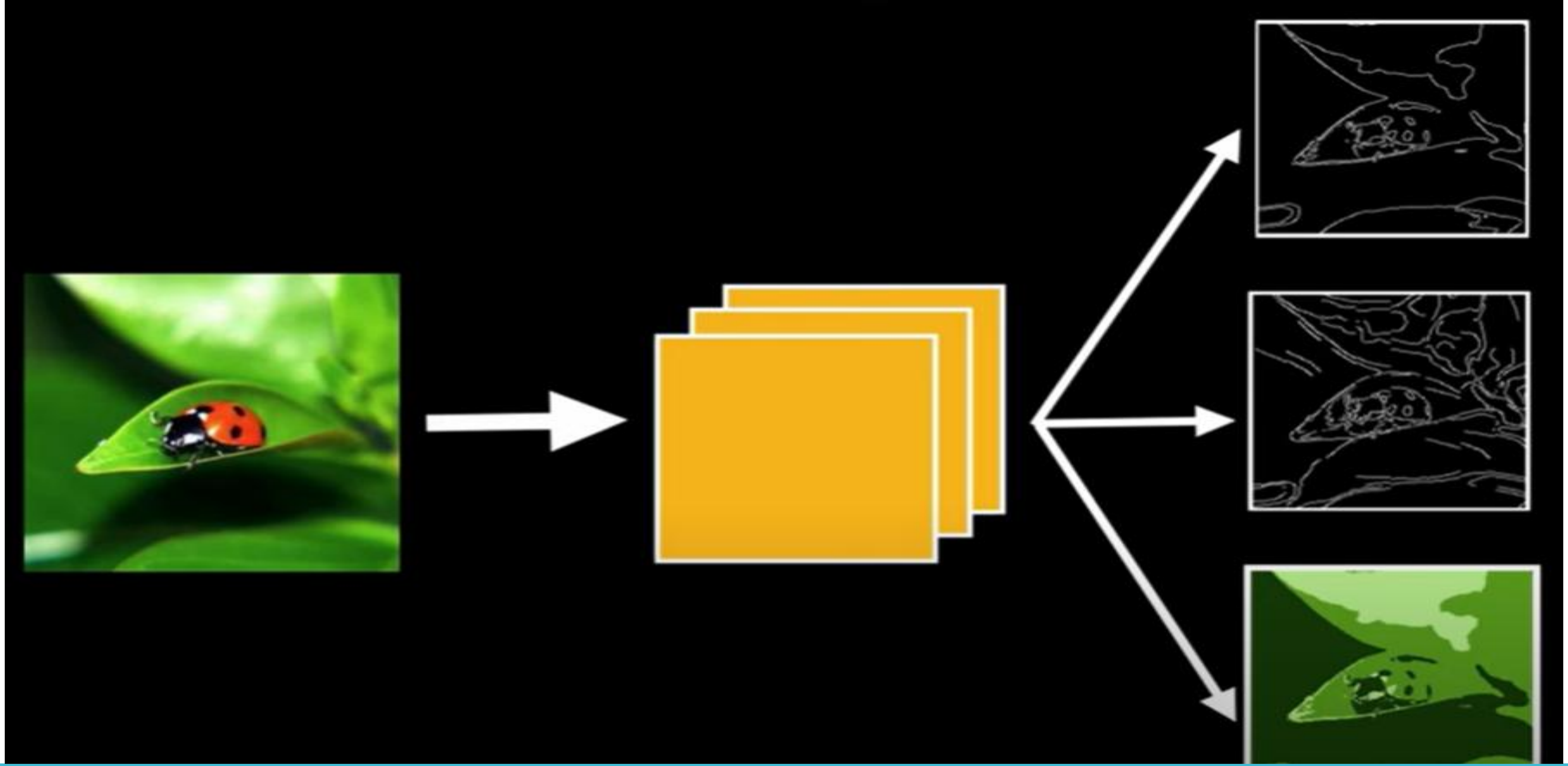


Love Apple phones
Self Attention

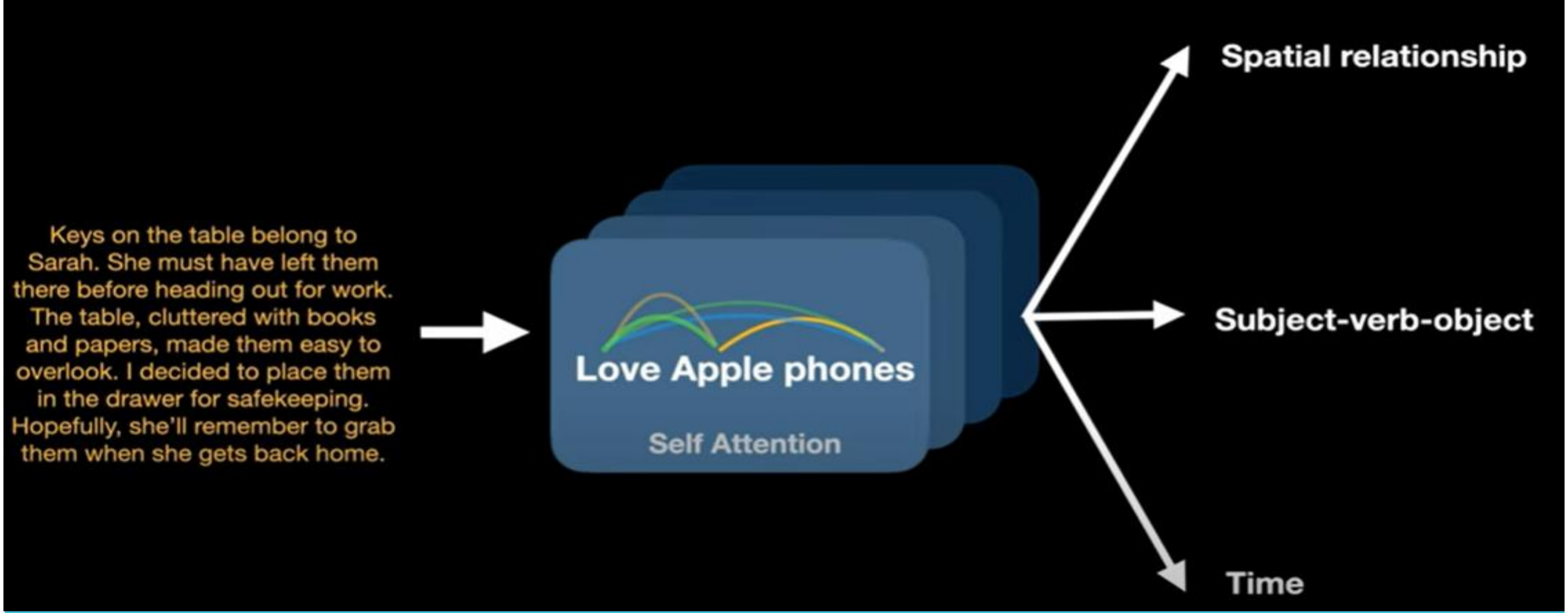
Multi-headed attention



Multi-headed attention

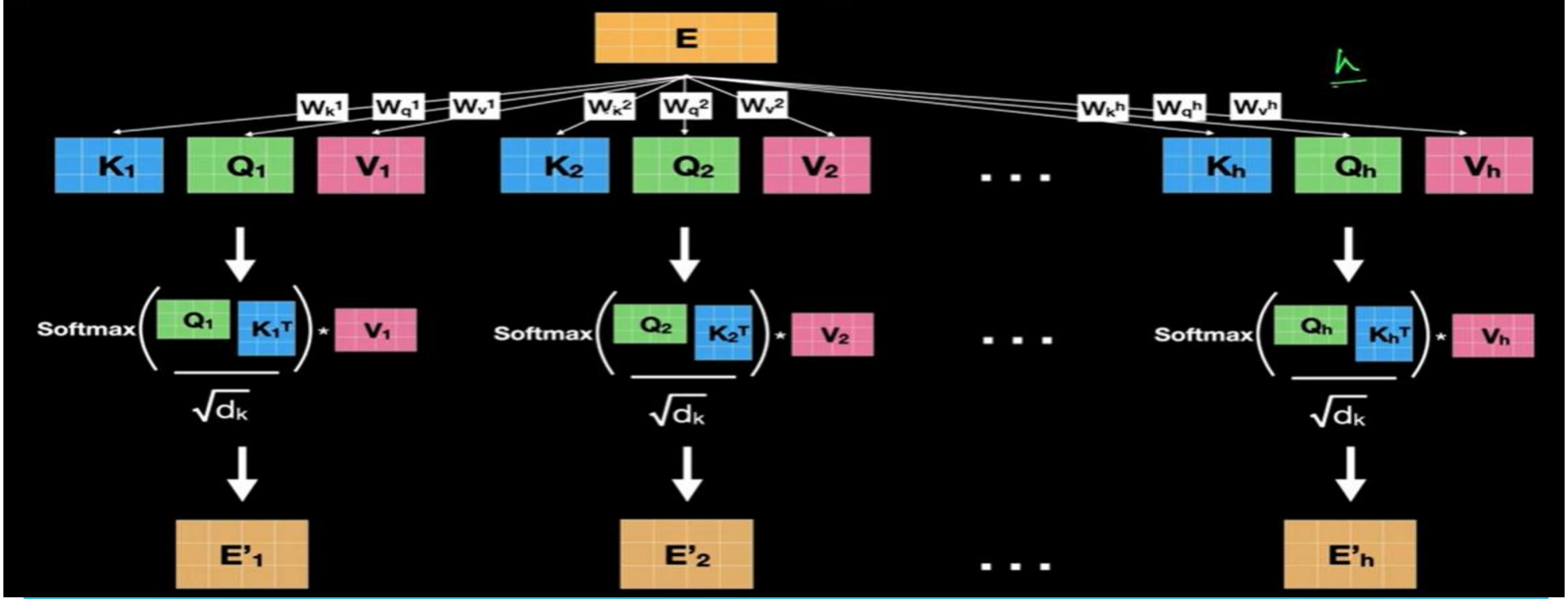


Multi-headed attention (in comparison to CNN)



Keys on the table belong to Sarah. She must have left them there before heading out for work. The table, cluttered with books and papers, made them easy to overlook. I decided to place them in the drawer for safekeeping. Hopefully, she'll remember to grab them when she gets back home.

Multi-headed attention (in comparison to CNN)


E
h
W_k^1
W_q^1
W_v^1
W_k^2
W_q^2
W_v^2
W_k^h
W_q^h
W_v^h
K_1
Q_1
V_1
K_2
Q_2
V_2
. . .
K_h
Q_h
V_h
Softmax$\left(\frac{Q_1 K_1^T}{\sqrt{d_k}}\right) * V_1$
Softmax$\left(\frac{Q_2 K_2^T}{\sqrt{d_k}}\right) * V_2$
. . .
Softmax$\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) * V_h$
E'$_1$
E'$_2$
. . .
E'$_h$

Multi-headed attention

1. Query (Q)

Represents the item that is **looking for relevant information**.

Example: A word currently being processed is the **query**, and it is "asking" which other words in the sequence it should pay attention to.

2. Key (K)

Represents **index-like information** about each item in the sequence.

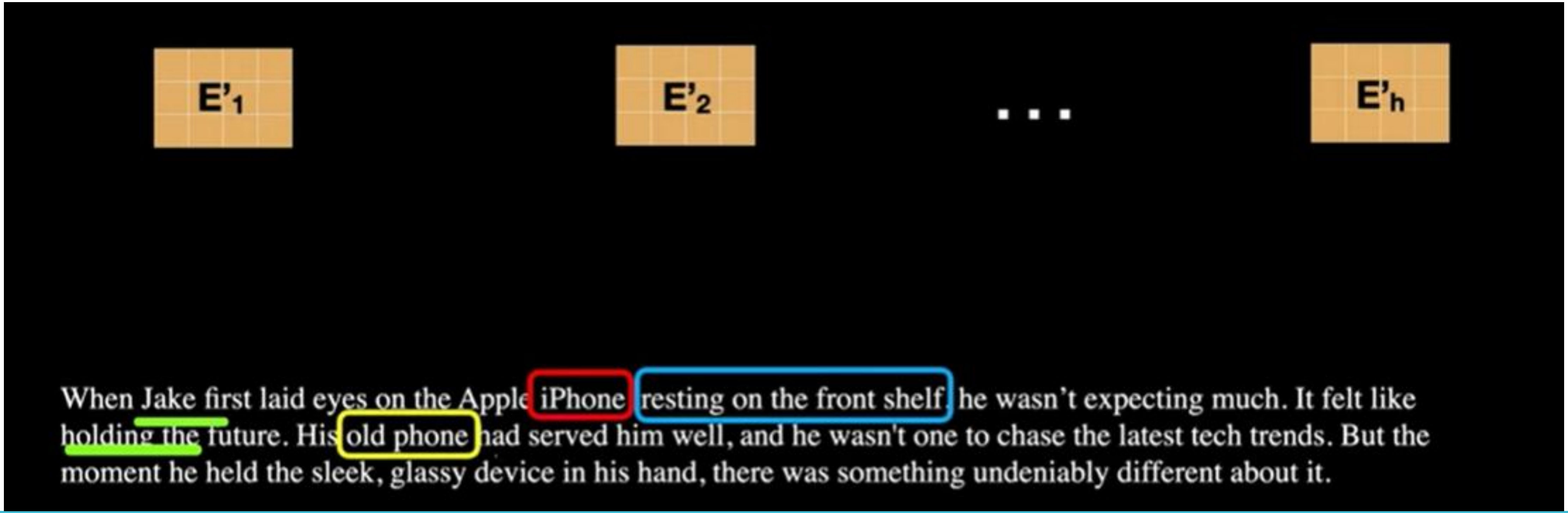
Each key vector is associated with a word in the input, and it helps determine **how relevant** that word is to the current query.

3. Value (V)

Contains the **actual information** to be retrieved if a key is deemed relevant.

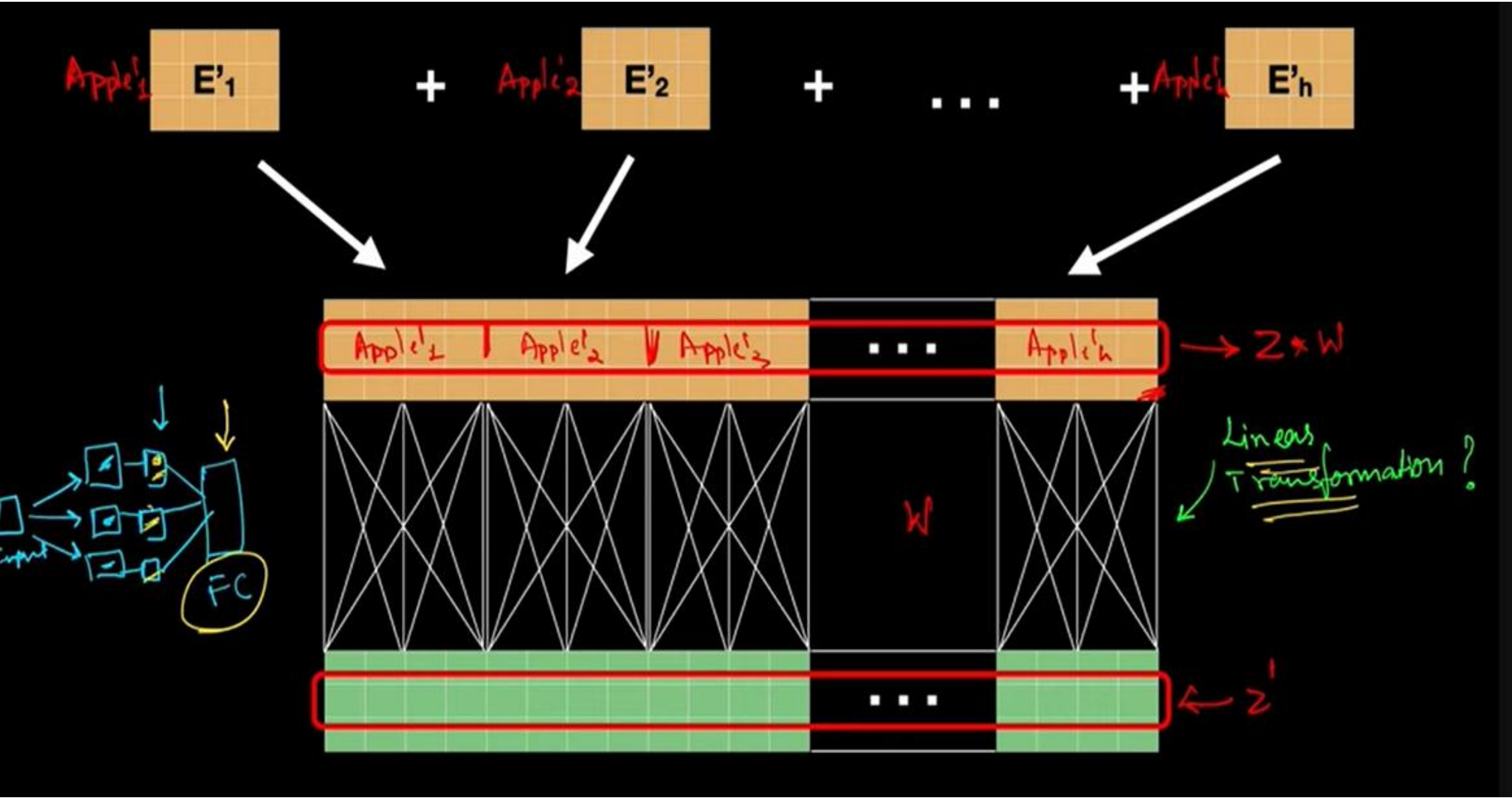
After identifying relevant keys (via matching with the query), the model **retrieves the values** corresponding to those keys.

Multi-headed attention (in comparison to CNN)



When Jake first laid eyes on the Apple iPhone resting on the front shelf he wasn’t expecting much. It felt like holding the future. His old phone had served him well, and he wasn't one to chase the latest tech trends. But the moment he held the sleek, glassy device in his hand, there was something undeniably different about it.

Multi-headed attention (in comparison to CNN)


Apple'1
E'1
+
Apple'2
E'2
+
...
+
Apple'h
E'h
Apple'1
Apple'2
Apple'3
...
Apple'h
→ Z * W
W
Linear Transformation ?
...
← Z'
Input
FC