

**Cross-Attention in Practice (Encoder-Decoder Architecture)**

In **encoder-decoder** architectures like the **Transformer** model, the encoder processes the source sequence and generates **key-value pairs**. These are then passed to the decoder during cross-attention.

1. **Encoder**:
   * The encoder takes the source sequence and produces **contextualized embeddings** for each token (e.g., words in a sentence).
   * The output from the encoder consists of key-value pairs (K,V)(K, V)(K,V) for each token.
2. **Decoder**:
   * The decoder takes the target sequence (e.g., an initial token or partially generated sequence) and generates queries based on the current target token.
   * The decoder uses the cross-attention mechanism to attend to the encoder's output, forming a context vector based on the most relevant tokens from the source sequence.
3. **Attention Flow**:
   * Cross-attention allows the decoder to focus on specific tokens in the source sequence that are most relevant to the current step in generating the target sequence.

For example, in **machine translation**:

* The encoder takes the input sentence in **English**: "I love cats."
* The decoder generates the target translation in **French**: "J'aime les chats."
* At each decoding step, the decoder will use cross-attention to focus on specific words in the English sentence ("I", "love", "cats") while generating the corresponding French words ("J'", "aime", "les").