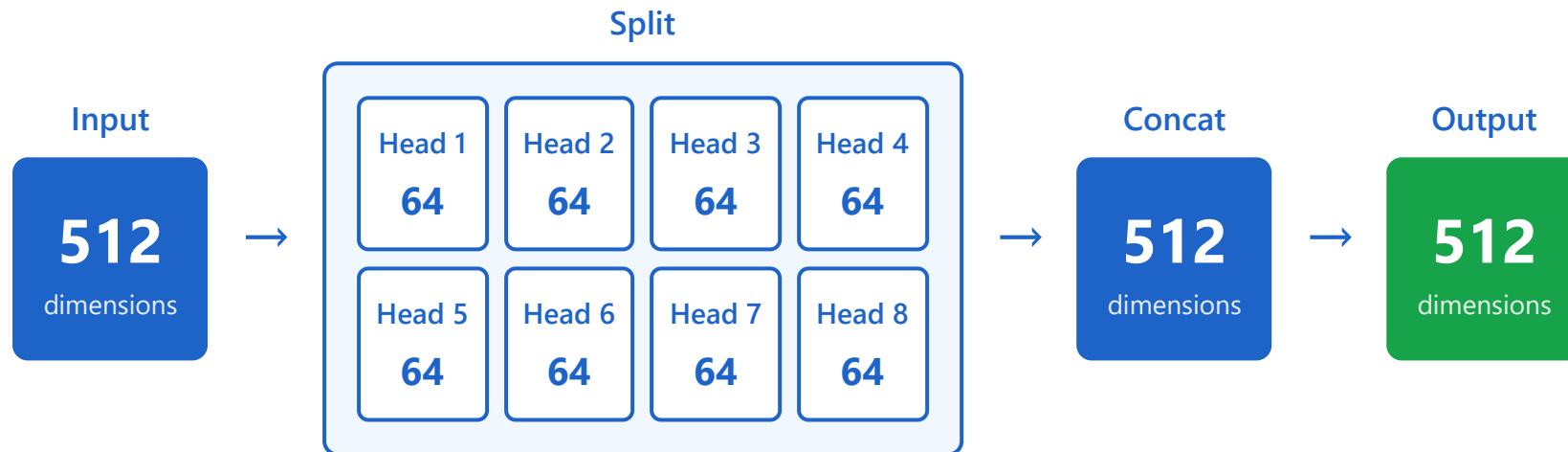


Multi-Head Operation Example

Configuration: `d_model = 512, h = 8` heads → Each head: `d_k = 64` dimensions



$8 \times 64 = 512$ | Each head learns different attention patterns **independently**



Each head learns
different patterns
independently



Computational cost
distributed
across heads



Parallelized
efficiently on
modern hardware

Detailed Calculation Example

Setup: $d_{\text{model}} = 4$, $h = 2$ heads, sequence length = 8 → Each head: $d_k = 2$ dimensions

Input Matrix X

Shape: (8, 4) - [sequence length × d_{model}]

```
token 1: [0.2, 0.5, 0.1, 0.8]
token 2: [0.3, 0.7, 0.4, 0.2]
token 3: [0.6, 0.1, 0.9, 0.5]
token 4: [0.4, 0.8, 0.3, 0.7]
token 5: [0.7, 0.2, 0.6, 0.4]
token 6: [0.5, 0.6, 0.2, 0.9]
token 7: [0.8, 0.3, 0.5, 0.1]
token 8: [0.1, 0.9, 0.7, 0.6]
```

Weight Matrices (per head)

W_Q - Query Weight

Shape: (4, 2)

W_K - Key Weight

Shape: (4, 2)

W_V - Value Weight

Shape: (4, 2)

Step-by-Step Calculation Process

Step 1: Split Input into 2 Heads

$X_{\text{head1}} = X[:, 0:2] \rightarrow \text{Shape: (8, 2)}$ - first 2 dimensions

$X_{\text{head2}} = X[:, 2:4] \rightarrow \text{Shape: (8, 2)}$ - last 2 dimensions

Step 2: Compute Q, K, V for Each Head

Head 1:

$Q_1 = X_{\text{head1}} \times W_{Q1} \rightarrow (8, 2) \times (2, 2) = (8, 2)$

$$K_1 = X_{\text{head}1} \times W_{K1} \rightarrow (8, 2) \times (2, 2) = (8, 2)$$

$$V_1 = X_{\text{head}1} \times W_{V1} \rightarrow (8, 2) \times (2, 2) = (8, 2)$$

Step 3: Calculate Attention Scores

$$\begin{aligned}\text{Scores}_1 &= (Q_1 \times K_1^T) / \sqrt{d_k} \\ &= (8, 2) \times (2, 8) / \sqrt{2} = (8, 8)\end{aligned}$$

Step 4: Apply Softmax & Compute Output

$$\text{Attention}_1 = \text{softmax}(\text{Scores}_1) \rightarrow (8, 8)$$

$$\text{Output}_1 = \text{Attention}_1 \times V_1 \rightarrow (8, 8) \times (8, 2) = (8, 2)$$

Step 5: Concatenate Head Outputs

$$\begin{aligned}\text{MultiHead} &= \text{Concat}(\text{Output}_1, \text{Output}_2) \\ &= \text{Concat}[(8, 2), (8, 2)] = (8, 4)\end{aligned}$$

Step 6: Final Linear Projection

$$\begin{aligned}\text{Final Output} &= \text{MultiHead} \times W_O \\ &= (8, 4) \times (4, 4) = (8, 4)\end{aligned}$$

Key Insight: Each head processes **2 dimensions** independently, learning different attention patterns. Final output maintains original **(8, 4)** shape.