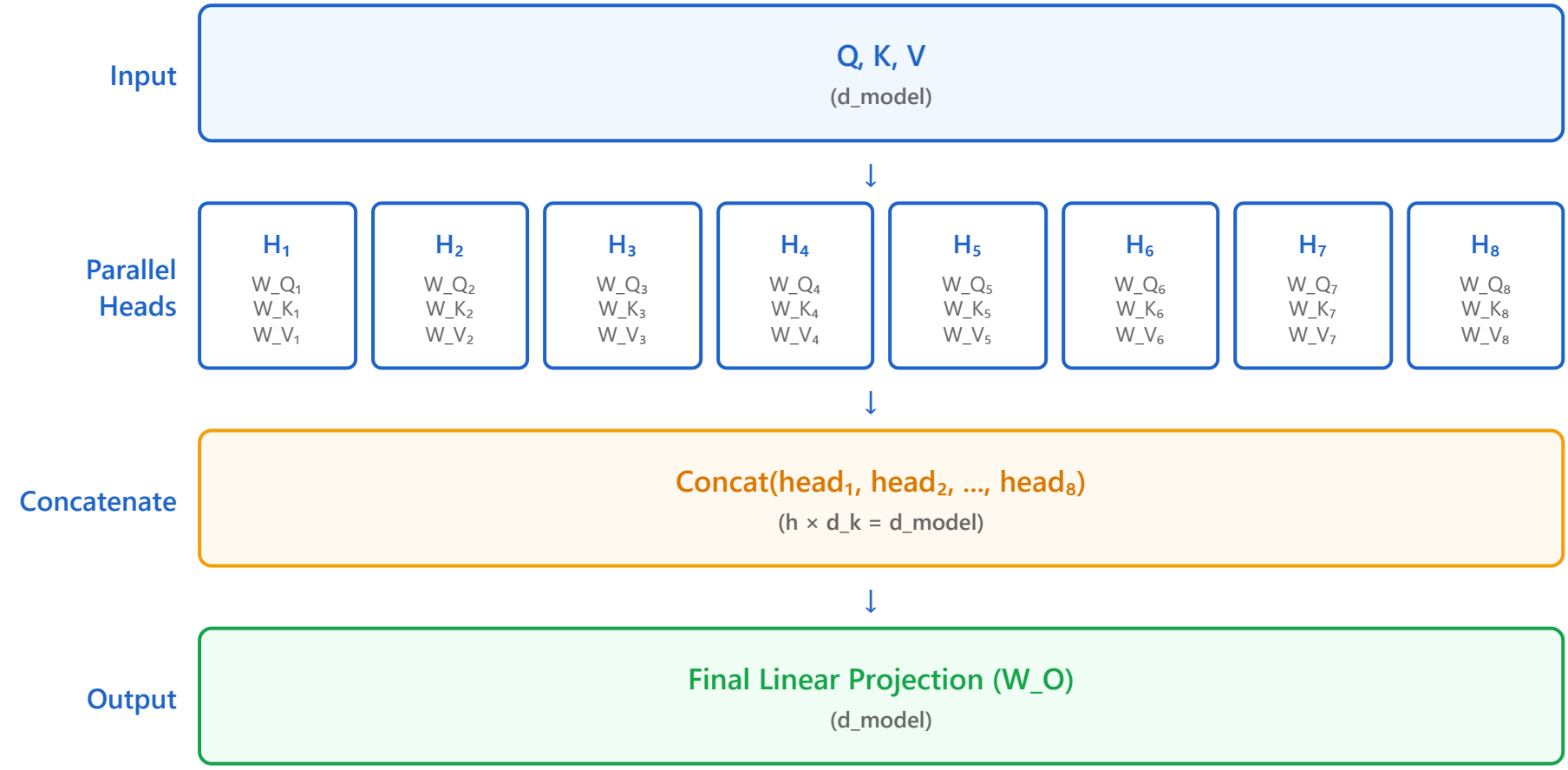
Multi-Head Attention Architecture



Configuration

Typically **h = 8** heads

Head dimension: **d_k = d_model / h**

Parameters

Each head: separate W_Q, W_K, W_V

Total params ≈ single large head

Formula

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$