Multi-Head Attention Implementation Points

Initialization & Configuration

Projection Matrices
Initialize with **appropriate scaling**

Head Dimensions
Use **same d_k** across all heads for simplicity

Regularization & Stability

Dropout
Apply after **attention weights** and **final output**

Layer Normalization
Typically after multi-head attention

Residual Connections
Help **gradient flow**

Optimization

Long Sequences
Consider **linear attention variants** for efficiency

✓ Efficient computation strategies

Best Practices

- Consistent head dimensions
- Proper normalization placement
- Efficient memory management

Key Trade-off

More heads increase model capacity but require **higher computational cost** — balance based on task requirements