Detailed Analysis of the Decoder

Single Decoder Layer (3 Sub-layers)

1. **Masked Multi-Head Self-Attention**
   Prevents attending to future positions

+ Residual → Layer Norm

2. **Encoder-Decoder Attention (Cross-Attention)**
   Q from decoder, K & V from encoder output

+ Residual → Layer Norm

3. **Feed-Forward Network**
   Position-wise transformation

+ Residual → Layer Norm

6 stacked layers with residual connections and layer normalization throughout

Masking
Prevents attending to future positions during training

Cross-Attention
Uses encoder output as Keys and Values

Residual Connections
Around all sub-layers for gradient flow

Layer Normalization
Applied throughout for stability

Comparison
Decoder has 3 sub-layers vs Encoder's 2 sub-layers