

## Model Parameters Distribution (14.8M total)

Memory Footprint:  
FP32: 59.9 MB  
FP16: 29.9 MB  
FP8: 14.9 MB

Vortex-Codec  
14,971,136 params  
(~14.8M)

Byte Embedding  
131,072 params  
(0.9%)

8 Transformer Blocks  
14,708,736 params  
(98.2%)

Output Projection  
131,328 params  
(0.9%)

### Per Transformer Block (1.84M params)

Layer Norms  
 $2 \times 512 = 1\text{K}$

Feed-Forward Network  
Linear  $512 \rightarrow 2048$ : 1,049K  
Linear  $2048 \rightarrow 512$ : 1,049K  
Total: 2,098K

Memory Compression  
Conv1D: 1,049K  
Layer norm: 0.5K  
Total: 1,049K

Multi-Head Attention  
Q, K, V projections: 786K  
Output projection: 262K  
Total: 1,048K