Multi-Head Attention

Comprehensive Demo and Analysis

Parallel attention heads with fused weight matrices,
reshape/transpose operations, and output projection.

Random seed: 42
Number of visualizations: 7

Generated by demo.py

Summary of Findings

1. Attention Patterns: Each head learns distinct attention distributions, visible as different heatmap structures across 8 heads.

2. Multi-Head vs Single-Head: Multiple heads provide richer representational diversity. Standard deviation across heads reveals complementary patterns.

3. Head Diversity: Pairwise cosine similarity between head attention matrices confirms heads attend to different positions/relationships.

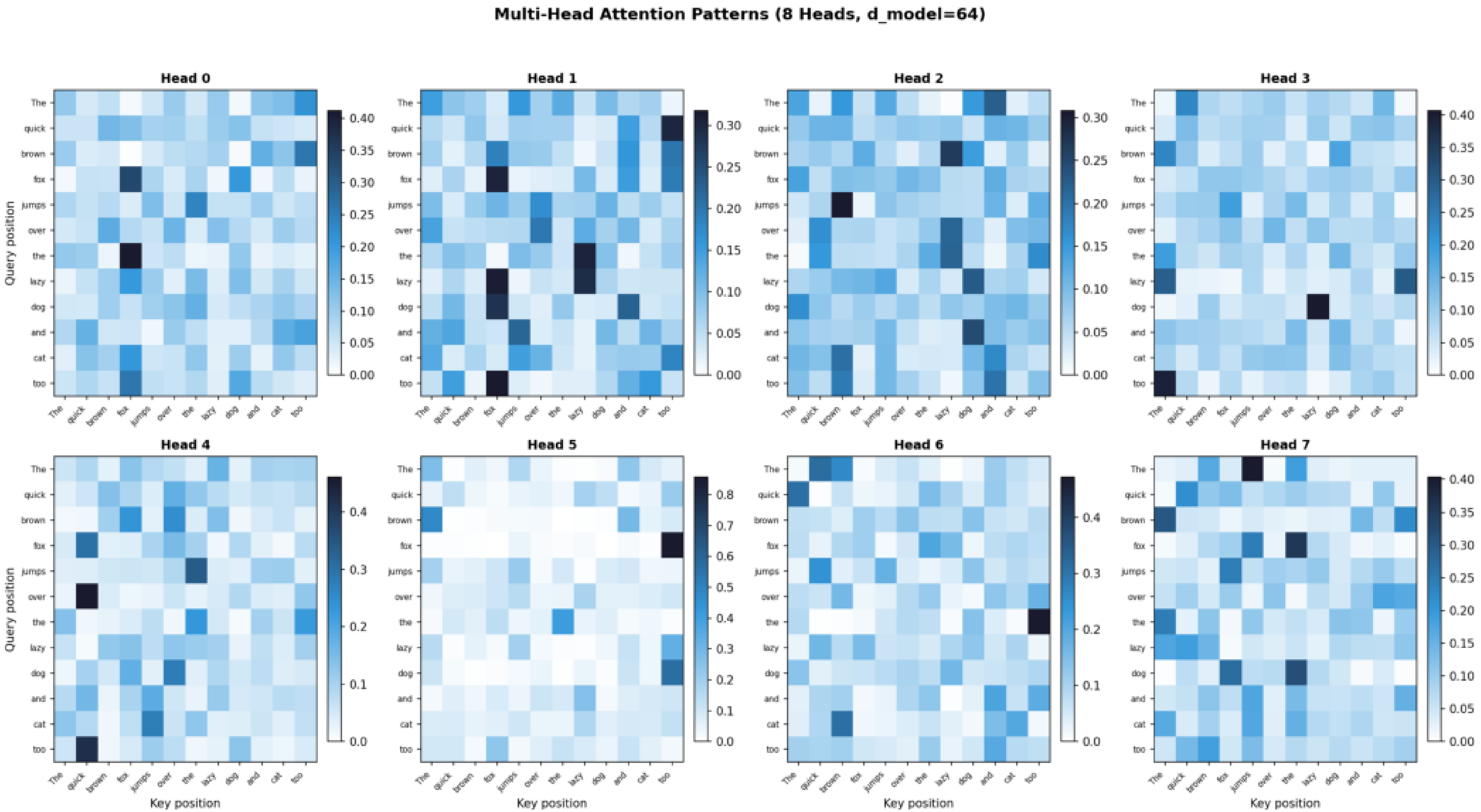
4. Causal Masking: Lower-triangular structure enforced correctly. Future positions receive zero attention weight. Row sums remain 1.0.

5. Memory Scaling: Attention matrix memory grows O(L^2). At long sequences (L>1024), the attention matrix dominates total intermediate memory.

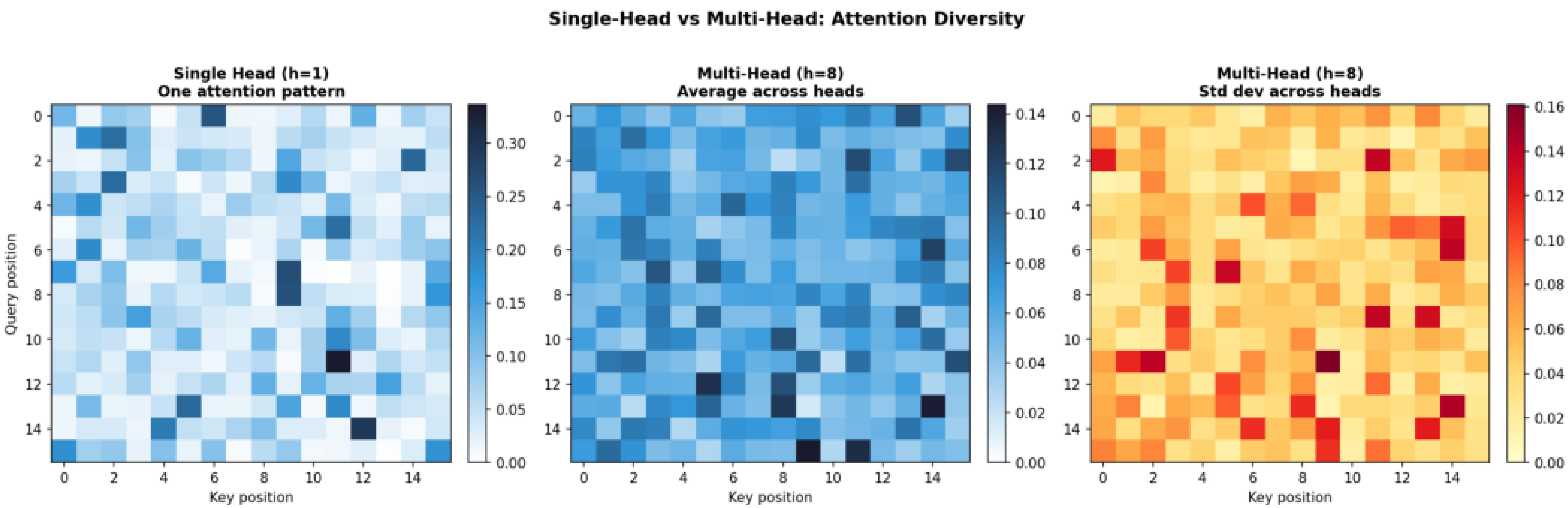
6. FLOPs Breakdown: Projection GEMMs dominate at short sequences. Attention core (QK^T + AV) overtakes at longer sequences.

7. Single-Head Equivalence: MHA(h=1) matches SelfAttention: PASS
   Both forward and backward pass agree to machine precision.

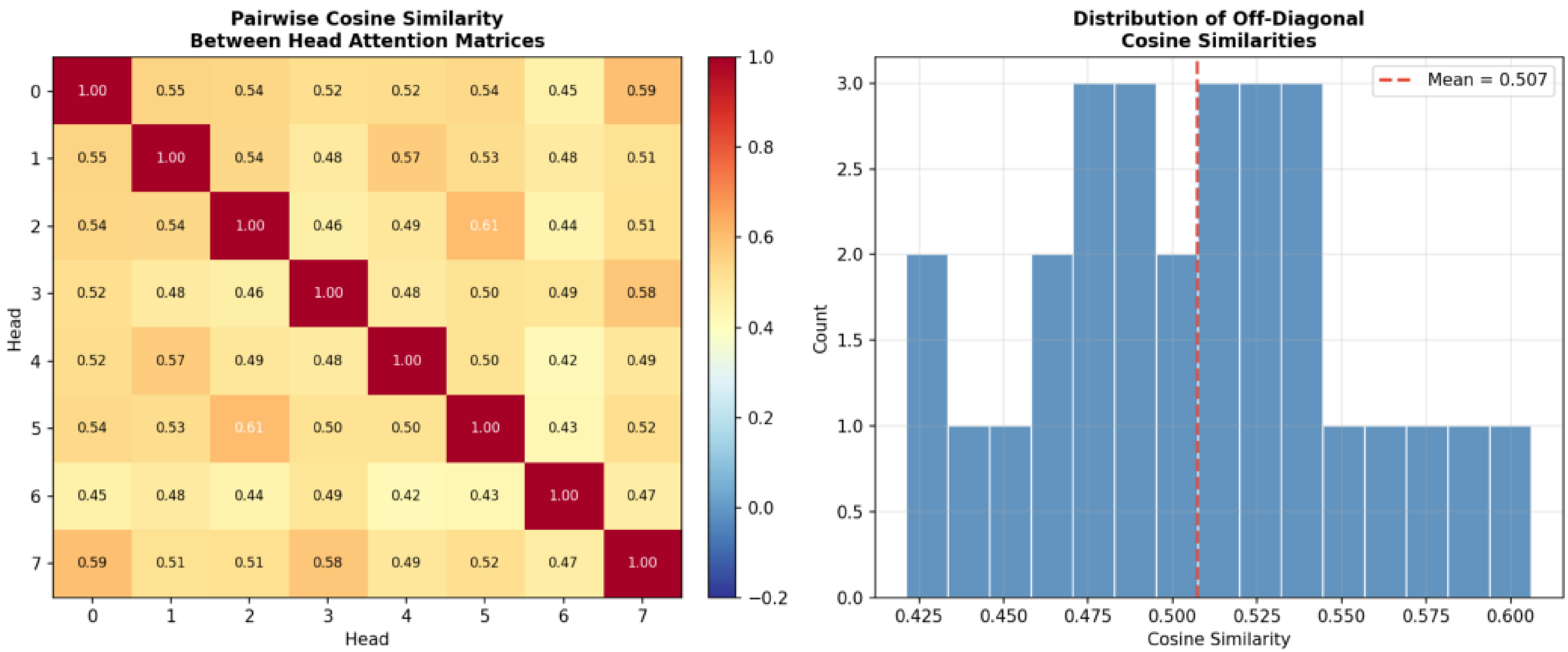
Example 1: Per-Head Attention Heatmaps



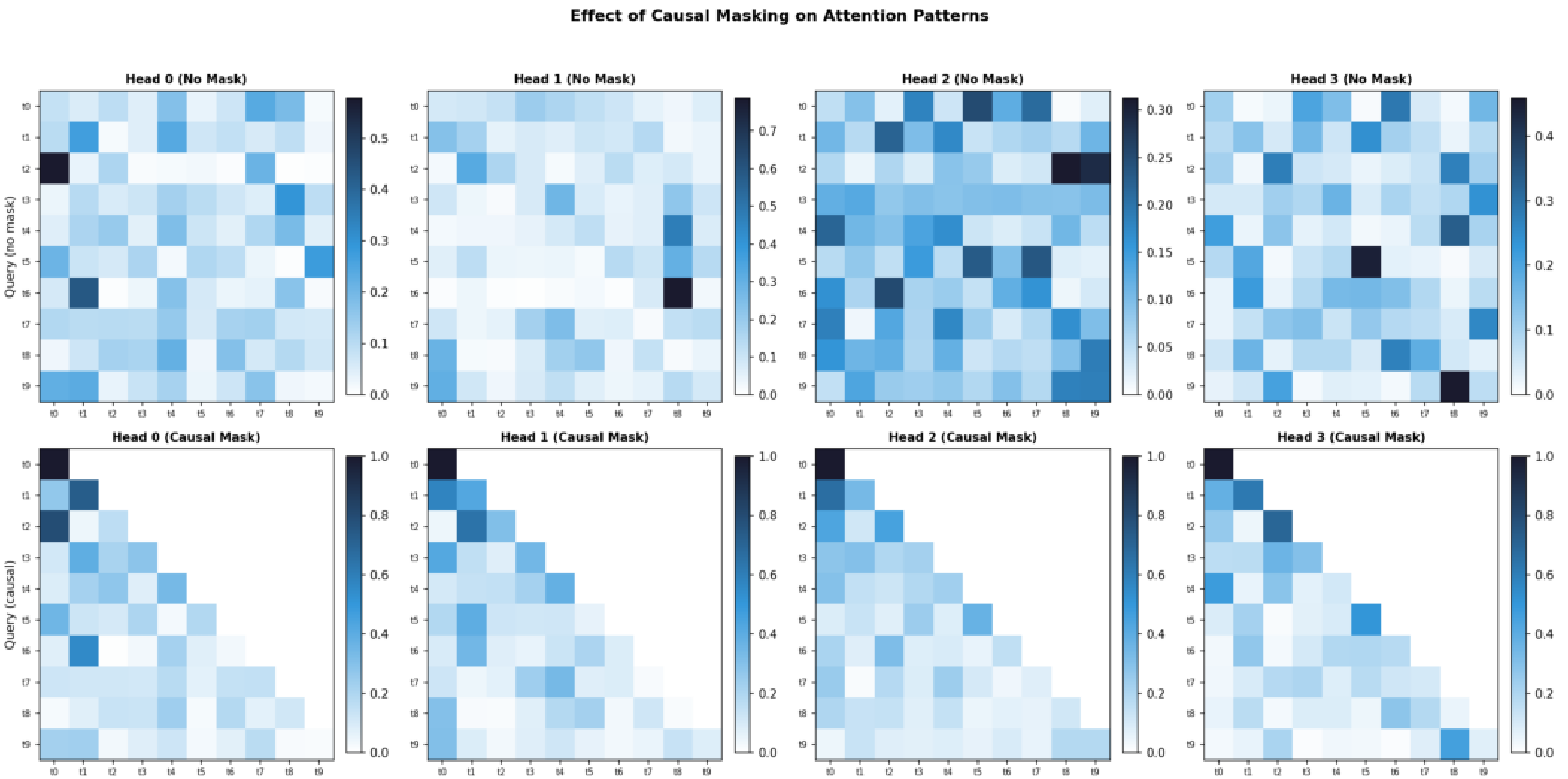
Example 2: Multi-Head vs Single-Head Comparison



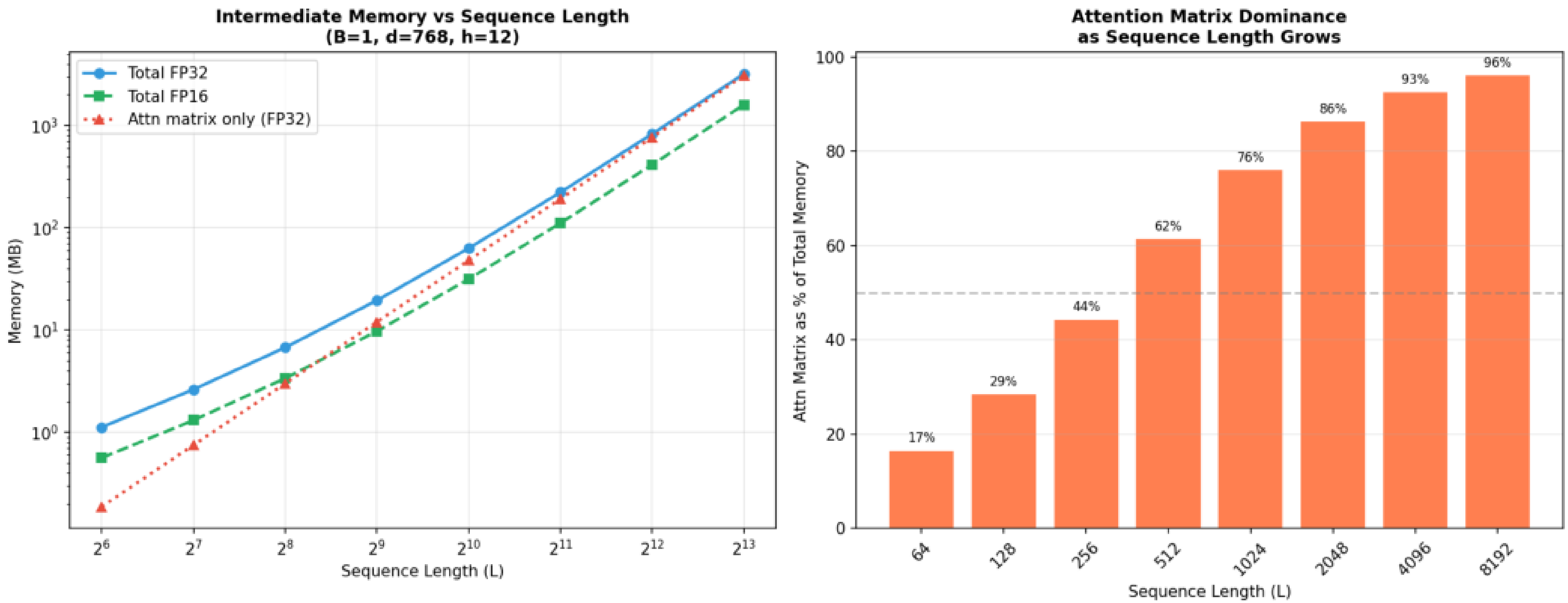
Example 3: Head Diversity Analysis (Cosine Similarity)



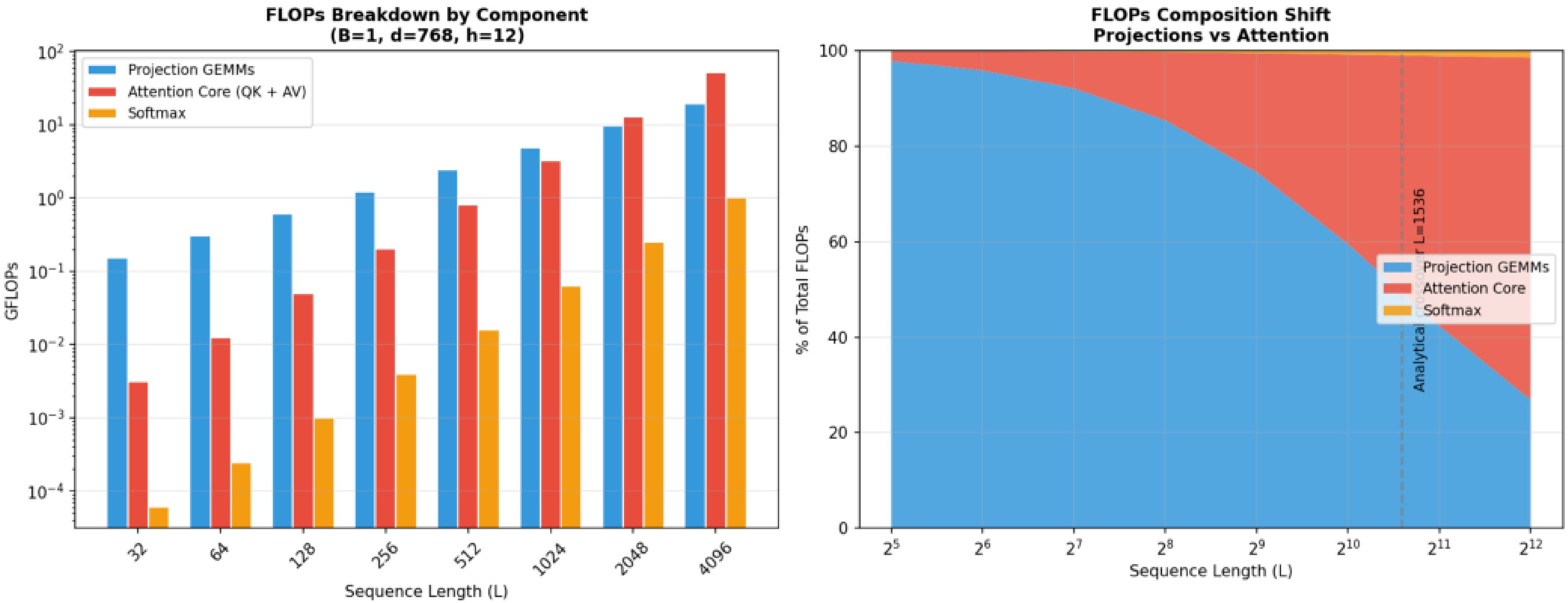
Example 4: Causal Masking Visualization



Example 5: Memory Scaling Analysis O(L^2)



Example 6: FLOPs Breakdown by Component



Example 7: Single-Head Equivalence Verification