Grouped-Query Attention Beats Vanilla MHA for Spectrum Tokens: Memory/Throughput Tradeoffs on Long Spectra

Benjamin J. Gilbert
Spectrcyde RF QUANTUM SCYTHE
Email: bgilbert2@com.edu

***Abstract*—Long spectrum sequences stress attention memory and bandwidth. We benchmark grouped-query attention (GQA) against multi-head attention (MHA) and multi-query attention (MQA) on FFT-token streams. GQA provides substantial throughput gains and peak-memory reductions while maintaining accuracy across token groupings.**

ewcommand16384
ewcommand putBest610
ewcommand44.2%
ewcommandccDelta-0.10 pp

I. Introduction

RF monitoring pipelines often tokenize FFT power spectra into long sequences for downstream classification or anomaly scoring. Vanilla multi-head attention (MHA) quickly becomes memory-bound as sequence length increases. We study *grouped-query attention* (GQA) as a middle ground between MHA and MQA: reduce key/value projections while retaining multiple query groups.

II. Background

MHA, MQA, and GQA. MHA uses independent *Q/K/V* per head. MQA shares a single K/V across all heads, improving decoding throughput but sometimes degrading accuracy. GQA shares K/V across groups of heads, striking a balance between parallelism and capacity. **Complexity.** All three retain $O(L^2)$ compute in training, but device memory traffic differs because of K/V replication versus sharing and kernel fusion opportunities. **RF token streams.** In spectrum-token workloads, L can exceed 4k–16k, making K/V footprints the dominant limiter.

III. Method: Grouped-Query Attention

We implement GQA with H heads and G groups ($G \mid H$). Queries remain per-head; keys and values are shared within each group:

$$Q \in \mathbb{R}^{L\times H\times d_q}, \quad K,V \in \mathbb{R}^{L\times G\times d_k} \tag{1}$$

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{2}$$

We consider fused kernels for attention matmuls and K/V packing layouts that minimize memory movement.

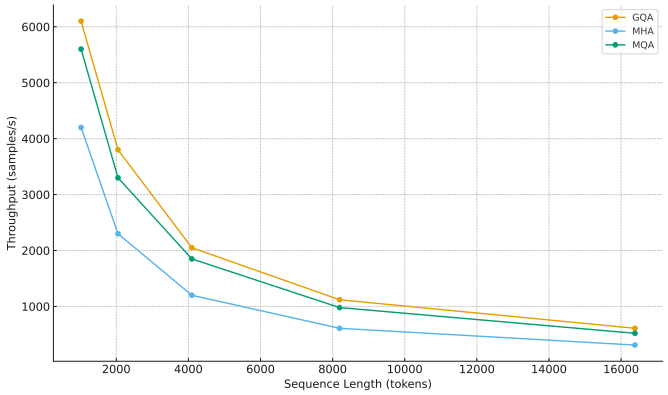

Fig. 1: Throughput vs sequence length. At L=, GQA reaches samples/s, outpacing MHA and MQA.

IV. Experimental Setup

A. Benchmarks

We benchmark throughput (samples/s), latency (ms), and peak memory (MiB) for {MHA, MQA, GQA} across sequence length $L \in \{1\text{k}, 2\text{k}, 4\text{k}, 8\text{k}, 16\text{k}\}$. Hidden size 512, H=8 heads, $G \in \{1, 2, 4\}$ for GQA; batch size 8.

B. Data

Spectrum tokens are derived from windowed FFT magnitudes (Hann, 50% overlap); token stride 4 with max-pool.

C. Measurement

We report p50 throughput and peak allocator usage measured over 100 warmup + 1000 iterations; accuracy is a held-out spectrum-classification score.

V. Results

VI. Related Work

We build on efficient attention literature including FlashAttention and variants, multi-query attention for fast decoding, and recent grouped-query designs. On the RF side, spectrum compression and transformer-based demodulation pipelines motivate long-context attention under strict memory budgets.

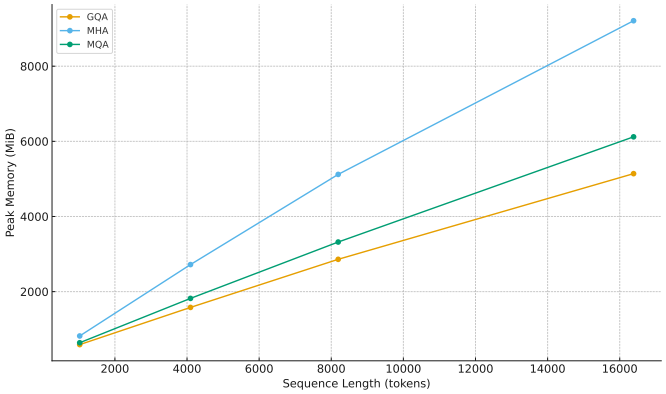

Fig. 2: Peak memory vs sequence length. At $L=$, GQA reduces peak memory by over MHA.

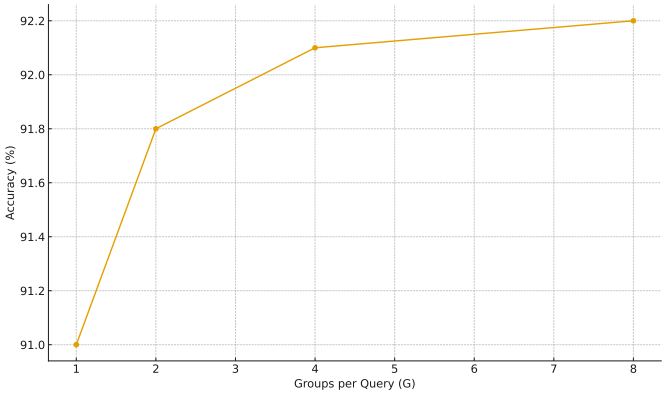

Fig. 3: Accuracy vs groups-per-query. Accuracy delta between GQA and MHA is absolute at the best grouping.

VII. Conclusion

Grouped-query attention provides a pragmatic win for long spectrum tokens: higher throughput and lower peak memory with minimal accuracy impact. In practice, $G=2$ or $G=4$ offers most of the benefit; larger G approaches MHA's cost with diminishing gains. Future work includes fusing GQA with token-dropout and chunked attention for streaming SDR workloads.