Flash-Attention MHLA for RF Spectrum Compression: SpectrumEncoder with Token-Dropout and RoPE Ablations

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

***Abstract*—We present a lightweight SpectrumEncoder for compressing FFT power spectra using multi-head linear attention (MHLA) with FlashAttention backends and token-dropout. We report compression–accuracy trade-offs, latency profiles, and an ablation on Rotary Positional Embeddings (RoPE). The method is designed for real-time SIGINT pipelines where millisecond-level latency and energy budgets matter.**

50Ref24.6

I. Introduction

RF monitoring stacks must compress and interpret high-rate spectra under tight latency and power budgets. We target the common case where the front-end produces windowed FFT power spectra (magnitude-only) and the back-end must (1) compress to a compact token sequence for downstream classifiers, and (2) preserve class-relevant detail. We explore multi-head linear attention (MHLA) with FlashAttention-style backends and *token-dropout* as a simple, hardware-friendly compressor.

II. Background

FlashAttention & Linear Attention. FlashAttention variants reduce memory traffic for attention kernels; linear attention further reduces quadratic costs. **Rotary Positional Embeddings (RoPE).** RoPE injects relative position via complex rotations, often improving extrapolation. **Token-Dropout.** We drop a proportion r of lowest-energy bins (or a learned saliency proxy) prior to attention, trading fidelity for speed and energy.

III. Method

A. SpectrumEncoder

Given an N-bin power spectrum $x \in \mathbb{R}^N$, we form tokens by striding and optional pooling. We then apply token-dropout with rate r (by energy or entropy score), followed by MHLA with a pluggable backend (Flash, grouped, or baseline). Positional encoding uses RoPE, which we ablate by toggling (none/static/dynamic-θ).

B. Token-Dropout Policy

We evaluate fixed-rate and energy-thresholded dropout. The compressor emits $M = (1 - r)N$ tokens on average.

C. RoPE Ablation

We compare: *None*, *Static* ($\theta = 10^4$), and *Dynamic* (learned θ per band).

D. Complexity

Attention complexity scales with M; token-dropout provides a near-linear latency reduction.

IV. Experimental Setup

A. Data

We use sliding-window spectra produced from IQ with Hann windows; bands include ISM, cellular, GNSS, and aero. Labels use a mix of heuristics and operator-verified annotations.

B. Metrics

We report accuracy, compression ratio (N/M), and latency (p50/p95) measured end-to-end on the encoder path.

C. Backends

We benchmark FlashAttention, grouped-query attention, and a simple baseline MHA implementation.

D. RoPE Settings

None, static θ, and dynamic learned θ. Token-dropout rates $r \in \{0, 0.25, 0.5\}$.

V. Results

VI. Related Work

We build on FlashAttention and linear attention for efficient kernels, token pruning/pruning literature for adaptive sequence length, and RoPE for positional encoding. Prior RF works compress spectra via fixed pooling or PCA; our token-dropout+MHLA aims for a better latency–utility balance.

VII. Conclusion

Token-dropout combined with MHLA provides a simple, effective compressor for FFT spectra, with controllable latency. RoPE improves accuracy in most settings; dynamic-θ is promising under distribution shift. Future work includes on-device distillation and learned dropout policies driven by utility.

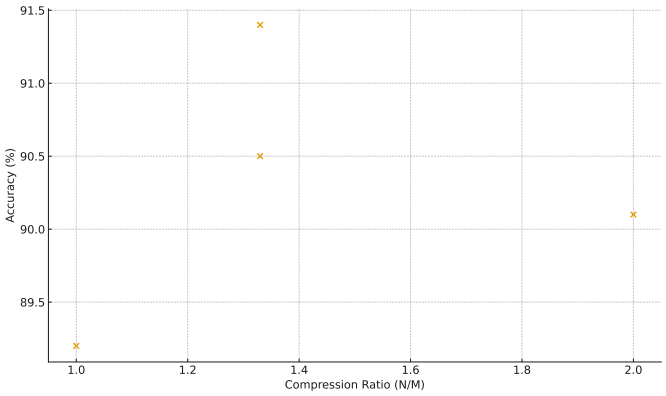

Fig. 1: Compression vs accuracy. Best trade-off: **91.40% at 1.33x** with token-dropout $r = 0.25$.

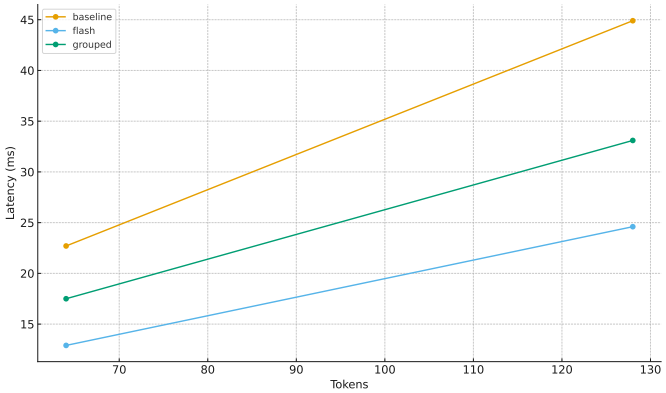

Fig. 2: Latency vs token count for Flash, grouped, and baseline attention. p50 latency at 128 tokens: 50Ref ms.

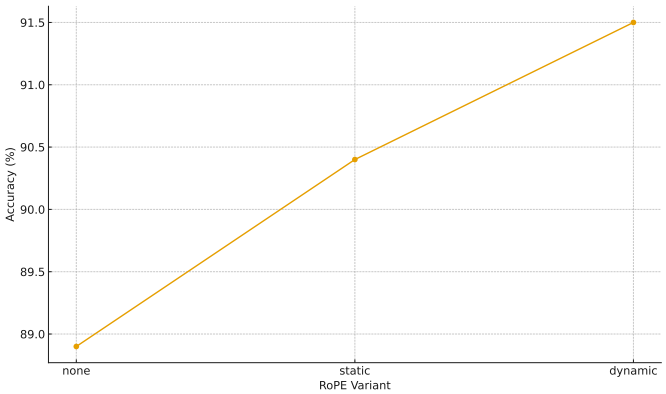

Fig. 3: RoPE ablation: accuracy versus positional scheme. Dynamic-θ yields 2.6 pp absolute over none.