Qwen2-1.5B stacked weights (d=65536, n=512)
Sketch-and-ridge-regression

FlashSketch (Ours)
Dense Gaussian (cuBLAS)
SJLT (GraSS Kernel)
SJLT (cuSPARSE)
Subsampled Fast Hadamard Transform

Sketch dimension (k)
● 64
■ 256
▲ 512
◆ 1024
✚ 2048
✖ 4096

1.83x speedup

Relative residual

Wall-clock total time (ms)