(a) GEMM — Time (ms) vs Matrix Size: 1024^3, 2048^3, 4096^3, 8192^3

(b) Batched GEMM — Time (ms) vs Batch Size: 2, 4, 8, 16

(c) Fused GEMMs — Time (ms) vs Matrix Size: 1024^2, 2048^2, 4096^2, 8192^2

(d) FlashAttn (d=128) — Time (ms) vs Seq Length: 128, 256, 512, 768, 1K, 1.5K, 2K, 4K

(e) FlashAttn (d=256) — Time (ms) vs Seq Length: 128, 256, 512, 768, 1K, 1.5K, 2K, 4K

(f) BlockSparse (d=64) — vs Seq Length: 128, 256, 512, 768, 1024

(g) BlockSparse (d=128) — vs Seq Length: 128, 256, 512, 768, 1024

Frameworks:
- AffineGraph
- CuBLAS
- Triton
- CUTLASS
- FlashAttention-2

(AffineGraph = Ours)