All data put in tables:

Implementation	Sequence Length (seq_len)	Embedding Dimension (embed_dim)	Execution Time (ms)	Relative Error
Naive Attention	1024	128	56.44	9.63E-07
Naive Attention	1024	256	112.20	7.72E-07
Naive Attention	1024	512	222.82	1.37E-06
Naive Attention	1024	1024	444.26	1.37E-06
Naive Attention	2048	256	228.81	1.1E-06
Naive Attention	2048	512	455.47	1.54E-06
Naive Attention	2048	1024	905.73	1.59E-06
Naive Attention	2048	2048	1806.39	2E-06
Naive Attention	4096	512	912.46	1.89E-06
Naive Attention	4096	1024	1820.77	1.58E-06
Naive Attention	4096	2048	3630.53	2.72E-06
Naive Attention	4096	4096	7266.81	2.7E-06

Implementation	Sequence Length (seq_len)	Embedding Dimension (embed_dim)	Execution Time (ms)	Relative Error
Fused Attention	1024	128	120.89	9.73E-07
Fused Attention	1024	256	242.67	7.69E-07
Fused Attention	1024	512	485.47	1.39E-06
Fused Attention	1024	1024	973.28	1.39E-06
Fused Attention	2048	256	485.31	1.13E-06
Fused Attention	2048	512	973.10	1.56E-06
Fused Attention	2048	1024	1947.50	1.63E-06
Fused Attention	2048	2048	3895.65	2.01E-06
Fused Attention	4096	512	1948.77	1.91E-06
Fused Attention	4096	1024	3894.31	1.61E-06
Fused Attention	4096	2048	7790.67	2.72E-06
Fused Attention	4096	4096	15579.84	2.7E-06

Implementation	Sequence Length (seq_len)	Embedding Dimension (embed_dim)	Execution Time (ms)	Relative Error
Tensor Core Fused Attention	1024	128	2.30	9.63E-07
Tensor Core Fused Attention	1024	256	3.22	7.72E-07
Tensor Core Fused Attention	1024	512	5.06	1.37E-06
Tensor Core Fused Attention	1024	1024	8.71	1.37E-06
Tensor Core Fused Attention	2048	256	9.91	1.1E-06
Tensor Core Fused Attention	2048	512	16.99	1.54E-06
Tensor Core Fused Attention	2048	1024	31.10	1.59E-06
Tensor Core Fused Attention	2048	2048	59.50	2E-06
Tensor Core Fused Attention	4096	512	61.78	1.89E-06
Tensor Core Fused Attention	4096	1024	120.95	1.58E-06
Tensor Core Fused Attention	4096	2048	269.52	2.72E-06
Tensor Core Fused Attention	4096	4096	531.71	2.7E-06

Implementation	Sequence Length (seq_len)	Embedding Dimension (embed_dim)	Execution Time (ms)	Relative Error
Block-Sparse Attention (Causal)	1024	128	110.51	1.55
Block-Sparse Attention (Causal)	1024	256	221.90	1.47
Block-Sparse Attention (Causal)	1024	512	446.83	1.51
Block-Sparse Attention (Causal)	1024	1024	890.62	1.52
Block-Sparse Attention (Causal)	2048	256	468.81	1.75
Block-Sparse Attention (Causal)	2048	512	937.78	1.67
Block-Sparse Attention (Causal)	2048	1024	1868.83	1.69
Block-Sparse Attention (Causal)	2048	2048	3739.90	1.70
Block-Sparse Attention (Causal)	4096	512	1916.61	1.87
Block-Sparse Attention (Causal)	4096	1024	3820.43	1.93
Block-Sparse Attention (Causal)	4096	2048	7633.66	1.89
Block-Sparse Attention (Causal)	4096	4096	1526.21	1.88

Implementation	Sequence Length (seq_len)	Embedding Dimension (embed_dim)	Execution Time (ms)	Relative Error
Block-Sparse Attention (Local)	1024	128	23.11	2.15
Block-Sparse Attention (Local)	1024	256	46.42	2.09
Block-Sparse Attention (Local)	1024	512	93.20	2.12
Block-Sparse Attention (Local)	1024	1024	185.33	2.13
Block-Sparse Attention (Local)	2048	256	46.77	3.17
Block-Sparse Attention (Local)	2048	512	93.08	3.07
Block-Sparse Attention (Local)	2048	1024	185.35	3.10
Block-Sparse Attention (Local)	2048	2048	370.32	3.10
Block-Sparse Attention (Local)	4096	512	93.32	4.42
Block-Sparse Attention (Local)	4096	1024	185.90	4.50
Block-Sparse Attention (Local)	4096	2048	371.57	4.47
Block-Sparse Attention (Local)	4096	4096	742.81	4.46