seq512-hd128-16h-32b | Q=512, K=512, head_dim=128, heads=16, batch=32

Kernel	Avg Forward Time (ms)	Relative to Triton Flash
scaled_dot_prod_attention	22.08	2.70x slower
vectorized_torch	29.53	3.62x slower
vectorized_torch_compiled	7.78	1.05x faster
flash_attention_triton	**8.17**	**1.00x**