

# Attention Optimization Speedup by Technique

(A100-80GB, head\_dim=128, batch\_size=8)

