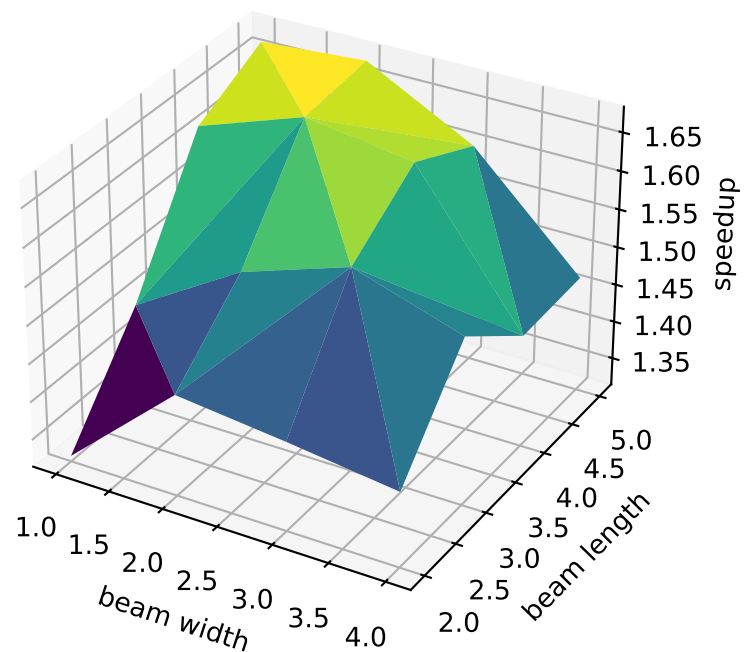
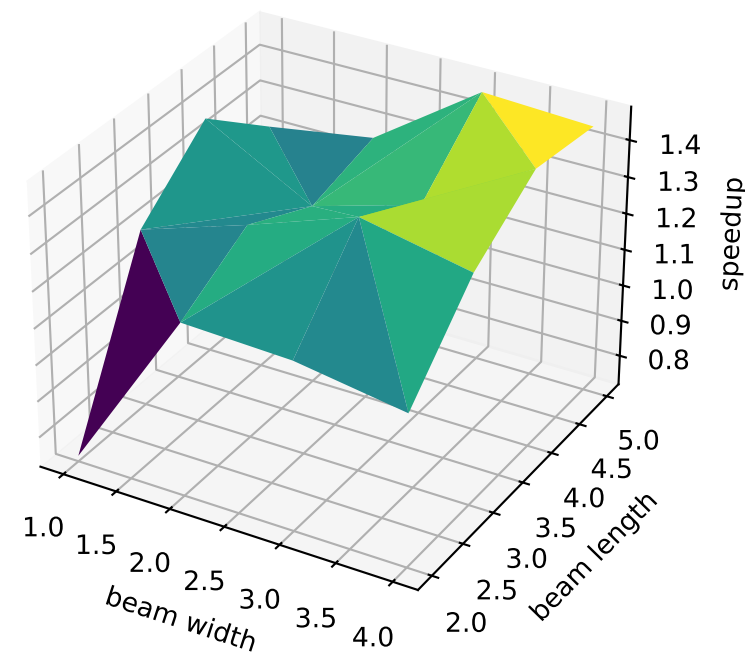


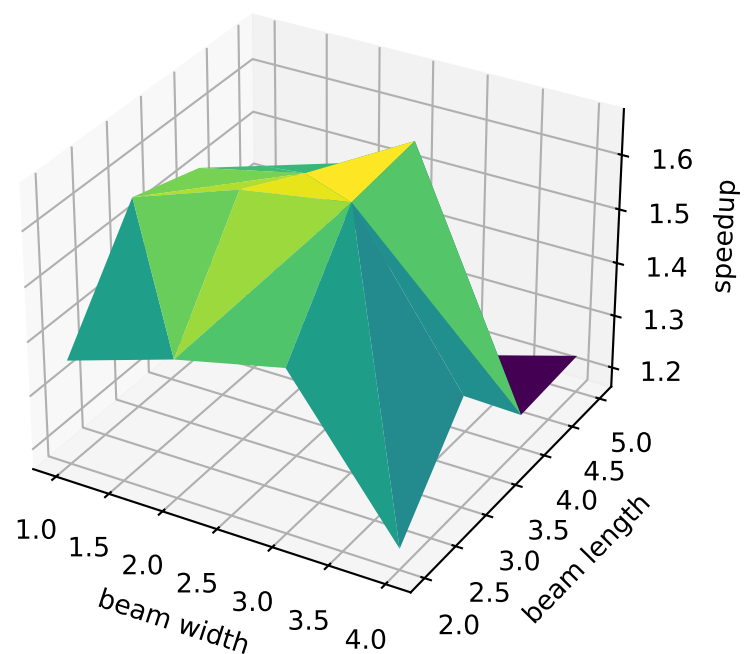
float16 greedy
beam shape=(2,5) 10.643 tokens/sec speedup=1.676



float16 non-greedy
beam shape=(4,4) 9.363 tokens/sec speedup=1.474



bfloat16 greedy
beam shape=(3,4) 10.580 tokens/sec speedup=1.672



bfloat16 non-greedy
beam shape=(4,3) 10.779 tokens/sec speedup=1.704

