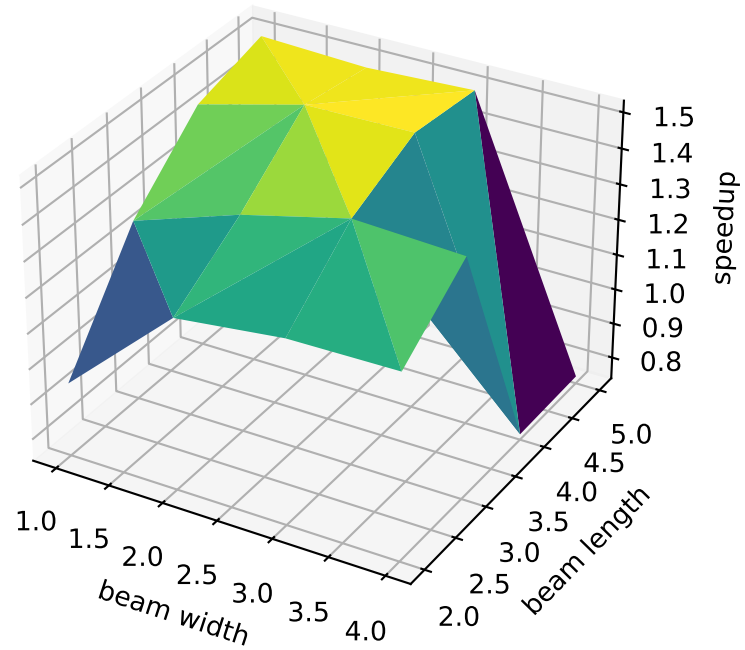
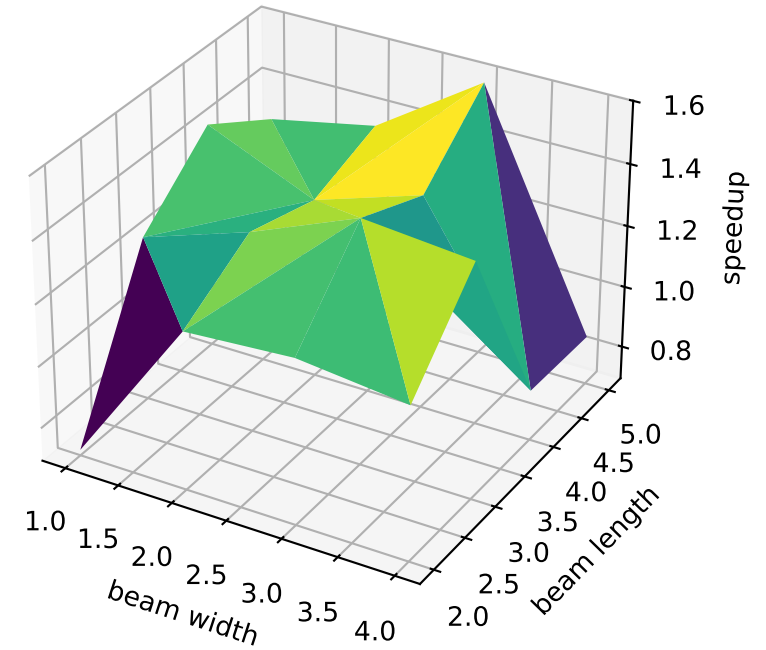


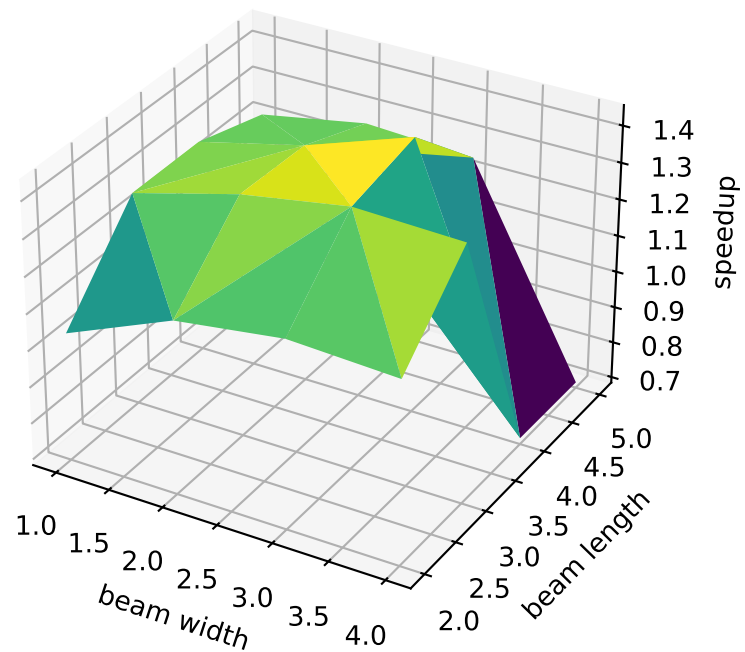
float16 greedy
beam shape=(3,4) 60.397 tokens/sec speedup=1.516



float16 non-greedy
beam shape=(3,5) 63.036 tokens/sec speedup=1.582



bfloat16 greedy
beam shape=(3,4) 55.697 tokens/sec speedup=1.439



bfloat16 non-greedy
beam shape=(4,3) 56.078 tokens/sec speedup=1.449

