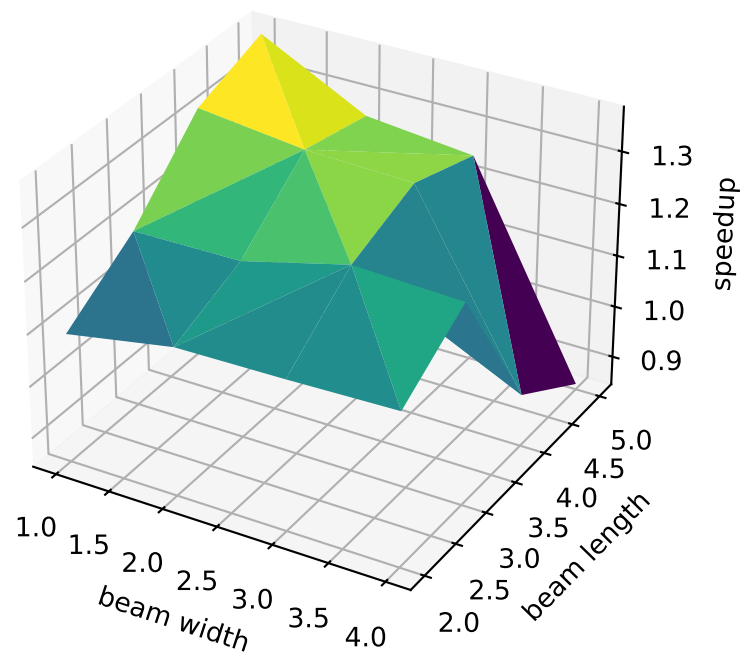
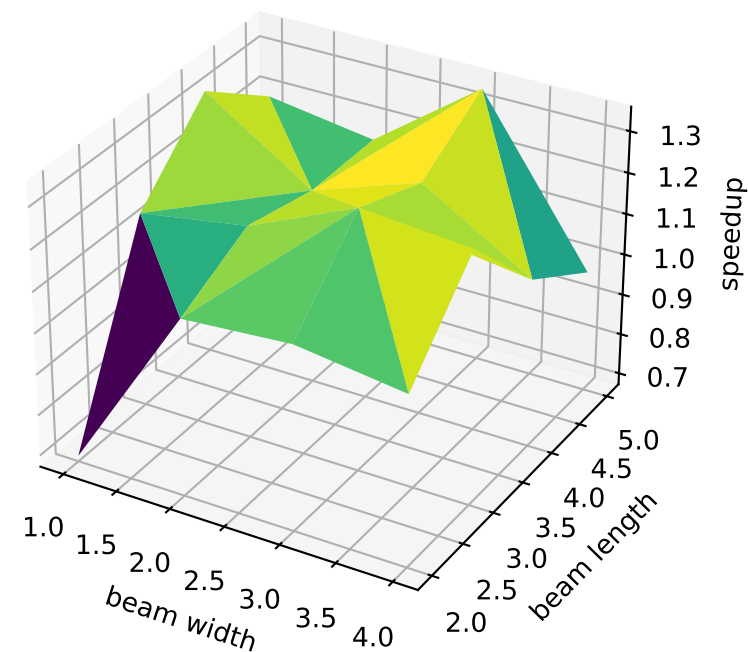


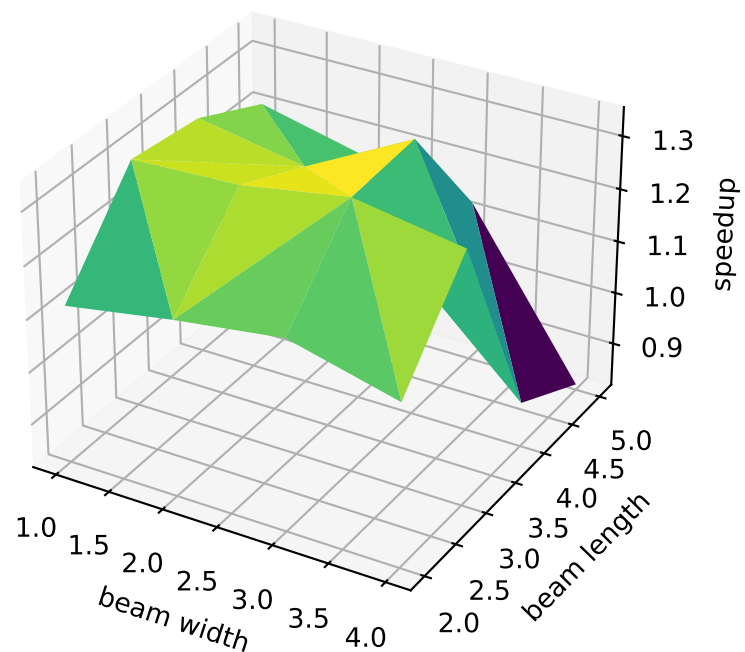
float16 greedy
beam shape=(1,5) 29.307 tokens/sec speedup=1.373



float16 non-greedy
beam shape=(3,5) 28.699 tokens/sec speedup=1.345



bfloat16 greedy
beam shape=(3,4) 27.723 tokens/sec speedup=1.343



bfloat16 non-greedy
beam shape=(4,3) 29.440 tokens/sec speedup=1.426

