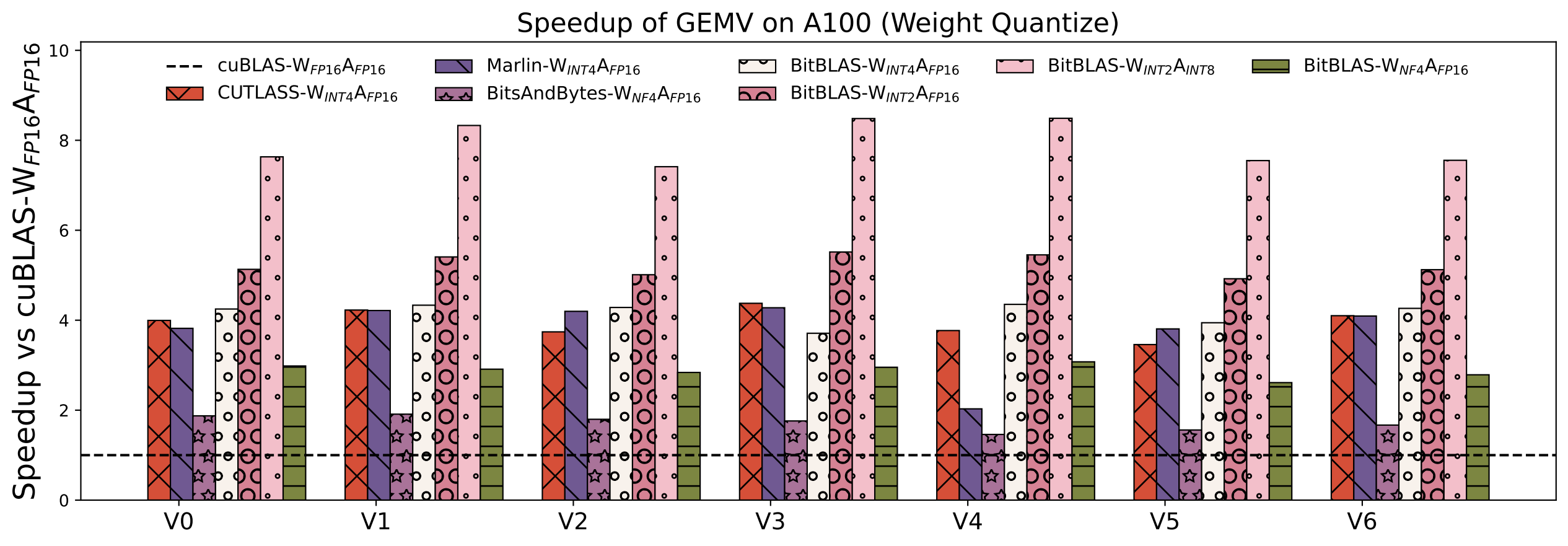
Speedup of GEMV on A100 (Weight Quantize)
Speedup vs cuBLAS-W$_{FP16}$A$_{FP16}$
cuBLAS-W$_{FP16}$A$_{FP16}$
CUTLASS-W$_{INT4}$A$_{FP16}$
Marlin-W$_{INT4}$A$_{FP16}$
BitsAndBytes-W$_{NF4}$A$_{FP16}$
BitBLAS-W$_{INT4}$A$_{FP16}$
BitBLAS-W$_{INT2}$A$_{FP16}$
BitBLAS-W$_{INT2}$A$_{INT8}$
BitBLAS-W$_{NF4}$A$_{FP16}$
0
2
4
6
8
10
V0
V1
V2
V3
V4
V5
V6